Qualitative Robustness in Bayesian Inference

Houman Owhadi and Clint Scovel
California Institute of Technology

December 10, 2014

Abstract

We develop a framework for quantifying the sensitivity of the distribution of posterior distributions with respect to perturbations of the prior and data generating distributions in the limit when the number of data points grows towards infinity. In this generalization of Hampel [47] and Cuevas' [18] notion of *qualitative robustness* to Bayesian inference, posterior distributions are analyzed as measure-valued random variables (measures randomized through the data) and their robustness is quantified using the total variation, Prokhorov, and Ky Fan metrics. Our results show that (1) the assumption that the prior has Kullback-Leibler support at the parameter value generating the data, classically used to prove consistency, can also be used to prove the non-robustness of posterior distributions with respect to infinitesimal perturbations (in total variation metric) of the class of priors satisfying that assumption, (2) for a prior which has global Kullback-Leibler support on a space which is not totally bounded, we can establish non *qualitative robustness* and (3) consistency and robustness are, to some degree, antagonistic requirements and a careful selection of the prior is important if both properties (or their approximations) are to be achieved. The mechanisms supporting our results are different and complementary to those discovered by Hampel and developed by Cuevas, and also indicate that misspecification generates non *qualitative robustness*.

Contents

1 Introduction

Robust Bayesian Analysis [6, 8] seeks to quantify variations in the output of Bayesian inference with respect to perturbations of the model and the prior. The importance of such an effort is expressed by Wasserman, Lavine and Wolpert [89]:

> "Most statisticians would acknowledge that an analysis is not complete unless the sensitivity of the conclusions to the assumptions is investigated. Yet, in practice, such sensitivity analyses are rarely used. This is because sensitivity analyses involve difficult computations that must often be tailored to the specific problem. This is especially true in Bayesian inference where the computations are already quite difficult."

In the classical Robust Bayesian framework, one defines a model and a prior and describes a perturbation class for the model and the prior and then asks, given the specific value of the data, what is the range of possible conclusions (posterior values) as the model and the prior varies over their perturbation classes. Although robustness has been established for classes of *finite dimensional* perturbations [6, 8, 12, 91], recent results [71, 70], summarized in [72], suggest that, under *finite information*, that is, *finite codimensional* perturbations (e.g., when the class of perturbed priors are defined via TV/Prokhorov metrics or a finite number of constraints on marginal distributions), Bayesian inference is not only not robust, it is brittle, that is, extremely non-robust.

The mechanism generating this brittleness can be seen as an extreme occurrence of the dilation phenomenon [90] which, in robust Bayesian inference, refers to the enlargement of optimal bounds caused by the data dependence of worst priors. This data dependence of worst priors is inherent to the formulation of classical Bayesian Sensitivity Analysis, in which worst priors are computed given the specific value of the data.

Therefore, although the brittleness results of [71, 70] suggest that Bayesian inference may not be robust under finite information within the classical Robust Bayesian framework [71, 70], one may ask whether robustness could be established under finite information by exiting this strict framework and computing the sensitivity of posterior conclusions independently of the specific value of the data. To investigate this question, this paper will generalize Hampel [47] and Cuevas' [18] notion of *qualitative robustness* to Bayesian inference based on the quantification of the sensitivity of the distribution of posterior distributions with respect to perturbations of the prior and the data generating distribution, in the limit when the number of data points grows towards infinity. Note that, contrary to classical Bayesian Sensitivity Analysis, in the proposed formulation, the data is not fixed and posterior values are therefore analyzed as dynamical systems randomized through the distribution of the data.

2 Structure of the paper and non-robustness mechanisms

2.1 General set up

We will now describe the general setup allowing us to analyze posterior distributions as *measure-valued random variables* or *measure-valued dynamical systems randomized through the distribution of the data*. We defer measure theoretic technicalities until later. Let Θ and X be measurable spaces. We write $\mathcal{M}(X)$ for the set of probability distributions on X, $\mathcal{M}(\Theta)$ for the set of probability distributions on Θ and $\mathcal{M}^2(\Theta) := \mathcal{M}\big(\mathcal{M}(\Theta)\big)$ the set of probability distributions on the set of probability distributions on Θ. We fix a prior $\pi \in \mathcal{M}(\Theta)$, a model $P : \Theta \to \mathcal{M}(X)$, and consider a data generating distribution $\mu \in \mathcal{M}(X)$. Let $\mu^n \in \mathcal{M}^n(X)$ denote the n-fold product measure corresponding to taking n-i.i.d. samples and for any such n-sample x^n from μ^n, let $\pi_{x^n} \in \mathcal{M}(\Theta)$ denote the posterior measure associated with the prior, model, and the sample. Then we ask how these posteriors π_{x^n} vary as a function of the sample data x^n when it is generated by i.i.d. sampling from μ, that is $x^n \sim \mu^n$. To do so, we consider the map

$$\bar{\pi} : X^n \to \mathcal{M}(\Theta)$$

defined by the determination of the posteriors

$$\bar{\pi}(x^n) := \pi_{x^n}$$

and consider its corresponding pushforward operator

$$\pi_* : \mathcal{M}(X^n) \to \mathcal{M}^2(\Theta)\,,$$

where we have removed the bar over π in the notation to remind us that this pushforward operator π_* corresponds to the prior π. Note that $\pi_*\mu^n$ is the sampling distribution of the posterior distribution π_{x^n} (of the parameter θ) when the data x^n is a random variable distributed according to μ^n. Since $\mathcal{M}^n(X) \subset \mathcal{M}(X^n)$ and $\mu^n \in \mathcal{M}^n(X)$, it follows that $\mu^n \in \mathcal{M}(X^n)$ so that the pushforward

$$\pi_*\mu^n \in \mathcal{M}^2(\Theta)$$

is well-defined.

In Section 4 we will define the *qualitative robustness* of Bayesian inference as the property that $\pi_*\mu^n$ and $\pi'_*(\mu')^n$ can be made arbitrarily close for large enough n if the priors π and π' and the data generating distributions μ and μ' are close enough. In particular, unlike Hampel and Cuevas who require "for all n" in their definitions, we follow Huber [50] and Mizera [66] in only requiring closeness "for large enough n". The results of this paper are applicable to both versions. Most of the non-robustness mechanisms presented in this paper only require perturbations in the prior distributions (that is, $\mu = \mu'$) and so are particularly relevant to well-specified Bayesian inference. Of course the notion of *qualitative robustness* will depend on the metrics placed on the space $\mathcal{M}(X)$ of data-generating distributions, the space $\mathcal{M}(\Theta)$ of prior distributions on Θ and the space $\mathcal{M}^2(\Theta)$ of random distributions on Θ. The total variation, Prokhorov and Ky Fan metrics will be of particular interest in our analysis.

2.2 Robustness and Consistency

Since the proposed notion of *qualitative robustness* is established in the limit when the number of data points grows towards infinity, it is natural to expect that the notion of *consistency* (i.e., the property that posterior distributions convergence towards the data generating distribution) will play an important role.

Although consistency is primarily a frequentist notion, according to Blackwell and Dubins [10] and Diaconis and Freedman [21], consistency is equivalent to *intersubjective agreement* which means that two Bayesians will ultimately have very close predictive distributions. Therefore, it also has importance for Bayesians. Fortunately, not only are there mild conditions which guarantee consistency, but the Bernstein-von-Mises theorem goes further in providing mild conditions under which the posterior is asymptotically normal. The most famous of these are Doob [23], Le Cam and Schwartz [62], and Schwartz [80, Thm. 6.1]. For more recent developments, see Barron, Schervish and Wasserman [3], Barron [4], Wasserman [88], Ghosh, Ghosal and Ramamoorthi [37], Ghosh [40], Kleijn [58], Walker [86], and the excellent review by Ghosal [36]. Moreover, the assumptions needed for this consistency are so mild that one can be lead to the conclusion that the prior does not really matter once there is enough data. For example, we quote Edwards, Lindeman and Savage [29]:

> "Frequently, the data so completely control your posterior opinion that there is no practical need to attend to the details of your prior opinion."

On the other hand, seemingly paradoxical results regarding inconsistency were found by Diaconis and Freedman [21, 31], and Freedman [30, 32, 33]: one consequence of Freedman [33] is that the set of pairs of priors leading to concurring asymptotic posterior conclusions is meager, i.e. very small in a topological sense.

Moreover, it appears that the full implications of model misspecification, when the data generating distribution is not exactly in the model class, have only been appreciated somewhat recently, see, e.g. Kleijn [57], Grünwald [44], Grünwald and Langford [43], Müller [67], Kleijn and van der Vaart [59, 60], and applied investigations in Biology begun in Douady et al. [24] and in Economics in Lubik and Schorfheide [63]. We quote from Kleijn and van der Vaart [60]:

> "In the misspecified situation the posterior distribution of a parameter shrinks to the point within the model at minimum Kullback-Leibler divergence to the true distribution, a consistency property that it shares with the maximum likelihood estimator. Consequently one can consider both the Bayesian procedure and the maximum likelihood estimator as estimates of this minimum Kullback- Leibler point. A confidence region for this minimum Kullback-Leibler point can be built around the maximum likelihood estimator based on its asymptotic normal distribution, involving the sandwich covariance. One might also hope that a Bayesian credible set automatically yields a valid confidence set for the minimum Kullback-Leibler point. However, the misspecified Bernstein-Von Mises theorem shows the latter to be false."

That is, in the misspecified case, not only do we have inconsistency, we also have strong convergence -with consequences for confidence sets. Indeed even earlier, to some, the consistency results appeared to generate more confidence than possibly they should. We quote A. W. F. Edwards [28, Pg. 60]:

> "It is sometimes said, in defence of the Bayesian concept, that the choice of prior distribution is unimportant in practice, because it hardly influences the posterior distribution at all when there are moderate amounts of data. The less said about this 'defence' the better."

We will demonstrate that the *Edwards defence* is essentially what produces non *qualitative robustness* in Bayesian inference. In particular, the assumptions required for consistency are such that arbitrarily small local perturbations of prior distributions (near the data generating distribution) result in consistency or non-consistency, and therefore, have large impacts on the asymptotic behavior of posterior distributions. To make this precise, in the following two sections we develop the notions of consistency and robustness of Bayesian inference. In particular, in Section 3 we develop the consistency theorem of Schwartz so that it can easily be used in the robustness analysis that is developed Section 4.

A relationship between *robustness and consistency* has been observed recently in Hable and Christmann [46] in ill-posed classification and regression problems, as defined by Dey and Ruymgaart [19], and is based on a fundamental link between robustness and

consistency in the results of Hampel [47, Lem. 3] and Cuevas [18, Thm. 1] which can be roughly stated as follows: qualitative robustness and consistency imply that the infinite sample limit is a continuous function of the data generating distribution. Therefore, for consistent systems non qualitative robustness follows from the discontinuity of the infinite sample limit. This approach has been used by Cuevas [18, Thm. 7] to establish the non qualitative robustness of two common Bayesian models with fixed priors under perturbations in the data generating distribution.

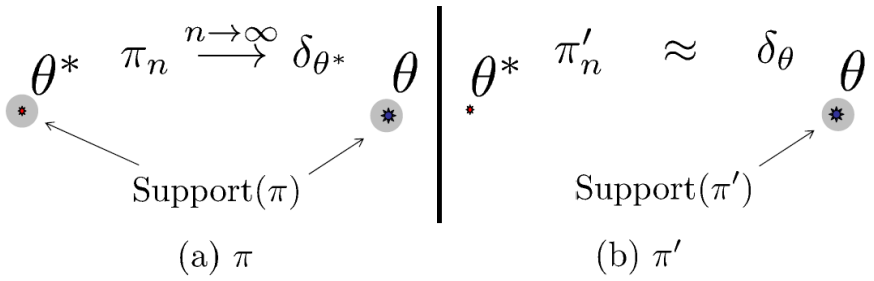
Moreover, with extra work and stronger but still mild assumptions, rates for consistency can be obtained, see e.g. Shen and Wasserman [81], Ghosal, Ghosh, and van der Vaart [39], Huang [49], and in the misspecified case Kleijn and van der Vaart [60]. In particular, the availability of *local* conditions guaranteeing rates to consistency has been investigated in Martin, Hong and Walker [65]. We conjecture that when these results are applicable, that they can be used to increase the degree of non *qualitative robustness* of Bayesian inference in a quantitative way.

Remark 2.1. Cuevas and Hampel use of the word *consistency* has a different meaning than the classical one used by, say, Schwartz. That is, to them, consistency denotes the "convergence" of the distribution of posterior values (in the infinite sample limit) towards a single point. This definition does not require the converge of the distribution of posterior distributions towards the correct point (i.e. the measure corresponding to the data generating distribution). Therefore, we could paraphrase Hampel and Cuevas' result as Convergence and Robustness implies continuity of the infinite sample limit with respect to the data generating distribution, whereas Schwartz proves consistency, that is convergence plus convergence to the correct point. Consequently, weaker theorems such as those found by Berk [9] in the misspecified case may be sufficient to interact with the result, or generalizations thereof, of Hampel and Cuevas to generate non *qualitative robustness*.

2.3 Mechanisms generating non-robustness

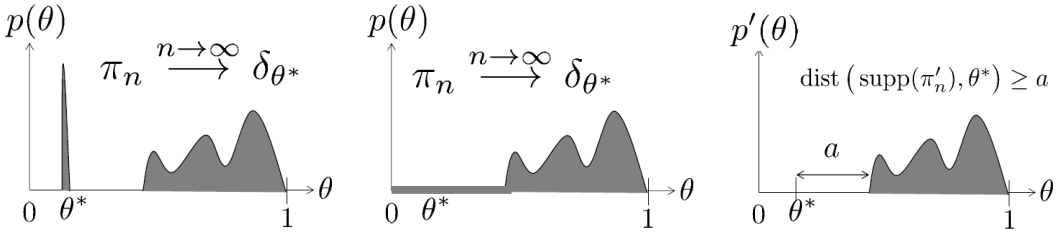
For the clarity of the paper, in this subsection, we introduce some of the mechanisms generating non *qualitative robustness* in Bayesian inference, which complement the mechanism discovered by Hampel [47, Lem. 3] and Cuevas [18, Thm. 1]. More precisely, the first illustrations presented are simplified graphical representations of the mechanisms used to obtain the main (non *qualitative robustness*) results of Section 5, which do not utilize any misspecification. Then we present mechanisms which appear to generate non qualitative robustness due to misspecification.

The core mechanism is derived from the nature of both the assumptions and assertions of results supporting consistency. A prototypical example of such results, which we will use in our proofs, is the corollary to Schwartz' consistency theorem presented in Section 3. More precisely, using the notations of Section 2.1, Corollary 3.1 states that if the data generating distribution is $\mu = P(\theta^*)$ and if the prior π attributes positive mass to every Kullback-Leibler neighborhood of $\theta^* \in \Theta$, then the posterior distribution converges towards δ_{θ^*} as $n \to \infty$. The assumption that π attributes positive mass to



(a) π (b) π'

Figure 2.1: The data generating distribution is $P(\theta^*)$. π' has most of its mass around θ. π is an arbitrarily small perturbation of π' so that π has Kullback-Leibler support at θ^*. Corollary 3.1 implies that π_n converges towards δ_{θ^*} while π'_n remains close to δ_θ.

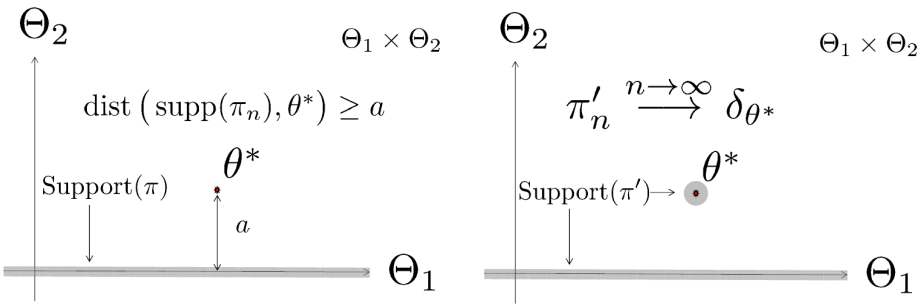


(a) Density of π (b) Density of π (c) Density of π'

Figure 2.2: The data generating distribution is $P(\theta^*)$. The probability density functions of π and π' are p and p' with respect to the uniform distribution on $[0, 1]$. π is an arbitrarily small perturbation of π' in total variation. π_n converges towards δ_{θ^*} while the distance between the support of π'_n and θ^* remains bounded from below by $a > 0$.

every Kullback-Leibler neighborhood of $\theta^* \in \Theta$ does not require π to place a significant amount of mass around θ^*, but instead can be satisfied with an arbitrarily small amount. Therefore, if, as in Figure 2.1, π is a prior distribution with support centered around $\theta \neq \theta^*$, but with a very small amount of mass about θ^*, so that it satisfies the assumptions of Corollary 3.1 at θ^*, then π can be slightly perturbed into a π' with support also centered around $\theta \neq \theta^*$, but with no mass about θ^*. In this situation, although π and π' can be made arbitrarily close in total variation distance, the posterior distribution of π converges towards δ_{θ^*} as $n \to \infty$, whereas that of π' remains close to δ_θ. Figure 2.2 gives an illustration of the same phenomenon when the parameter space Θ is the interval $[0, 1]$ and the probability density functions of π and π' are p and p' with respect to the uniform measure.

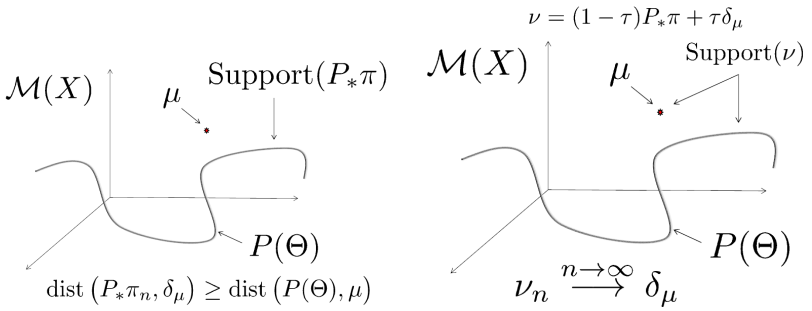
Note that the mechanism illustrated in Figures 2.1 and 2.2 does not generate non qualitative robustness at *all priors* but instead for the full class of *consistency priors*, de-



(a) π (b) π'

Figure 2.3: Non-robustness caused by misspecification. The parameter space of the model is Θ_1. We assume that the model $P : \Theta_1 \to \mathcal{M}(X)$ is the restriction of an injective model $\bar{P} : \Theta_1 \times \Theta_2 \to \mathcal{M}(X)$ to $\Theta_1 \times \{\theta_2 = 0\}$. The data generating distribution is $\bar{P}(\theta^*)$ where $\theta^* := (\theta_1^*, \theta_2^*)$, with $\theta_2^* \neq 0$, so that the model P is misspecified. π satisfies Cromwell's rule. π' is an arbitrarily small perturbation of π having Kullback-Leibler support at θ^*. Corollary 3.1 implies that π'_n converges towards δ_{θ^*} while the distance between the support of π_n and θ^* remains bounded from below by $a > 0$.

fined by the assumption of having positive mass on every Kullback-Leibler neighborhood of θ^*. One may wonder whether this non *qualitative robustness* can be avoided by selecting the prior π to satisfy Cromwell's rule (that is, the assumption that π gives strictly positive mass to every nontrivial open subset of the parameter space Θ). Theorem 5.4 shows that this is not the case if the parameter space Θ is not totally bounded. For example, when $\Theta = \mathbb{R}$, for all $\delta > 0$ one can find $\theta \in \mathbb{R}$ such that the mass that π places on the ball of center θ and radius one is smaller than δ, and by displacing this small amount of mass one obtains a perturbed prior π' whose posterior distribution remains asymptotically bounded away from that of π when the data-generating distribution is $P(\theta)$. Similarly if Θ is totally bounded then Theorem 5.5 places an upper bound on the size of the perturbation of the prior π that would be required as a function of the covering complexity of Θ. Note that these observations suggest that a maximally *qualitatively robust* prior should place as much mass as possible near all possible candidates θ for the parameter θ^* of the data generating distribution, thereby reinforcing the notion that a maximally robust prior should have its mass spread as uniformly as possible over the parameter space. We refer to Section 6 for a discussion on the existence of such measures in relation to the geometry of the sets of measures having local and global Kullback-Leibler support.



(a) $P_*\pi$ (b) ν

Figure 2.4: Non-robustness caused by misspecification. The parameter space of the model is Θ. The data generating distribution is $\mu \notin P(\Theta)$, so that the model is misspecified. ν is an arbitrarily small perturbation of $P_*\pi$ in total variation distance having non-zero mass on μ. Note that if a small mass on μ is sufficient to ensure the consistency of ν then such a mechanism also implies the non-robustness of the model with respect to misspecification since in such a case, ν_n, the posterior distribution of ν would converge towards δ_μ whereas the distance between the support of $P_*\pi_n$ and μ remains bounded from below by the distance from the model to the data generating distribution.

2.3.1 Robustness under misspecification

Although the main results of Section 5 do not utilize any model misspecification, the brittleness results of [71] suggest that misspecification should also generate non qualitative robustness. Indeed, although, one may find a prior that is both consistent and *qualitatively robust* when Θ is totally bounded and the model is well-specified, we now show how extensions of the mechanism illustrated in Figures 2.1 and 2.2 suggest that misspecification implies non *qualitative robustness*. Consider the example illustrated in Figure 2.3. In this example the model P is the restriction of a well specified larger model $\bar{P} : \Theta_1 \times \Theta_2 \to \mathcal{M}(X)$ to $\theta_2 = 0$. Assume that the data generating distribution is $\bar{P}(\theta_1^*, \theta_2^*)$ where $\theta_2^* \neq 0$, so that the restricted model P is misspecified. Let π be any prior distribution on $\Theta_1 \times \{\theta_2 = 0\}$. Although π may satisfy Cromwell's rule the mechanisms presented in this paper suggest that is not *qualitatively robust* with respect to perturbed priors having support on $\Theta_1 \times \Theta_2$. Indeed, let π' be an arbitrarily small perturbation of π obtained by removing some mass from the support of π and adding that mass around θ^*. Note that π' can be chosen arbitrarily close to π while satisfying the local consistency assumption of Corollary 3.1, which implies that the posterior distributions of π' concentrate on θ^* while the posterior distributions of π remain supported on $\Theta_1 \times \{\theta_2 = 0\}$. Note that if $\bar{P}$ is interpreted as an extension of the model P, then this mechanism suggests that we can establish conditions under which Bayesian inference is not *qualitatively robust under model extension*.

Figure 2.4 represents a non-parametric generalization of the mechanism of Figure 2.3. Assume that the data generating distribution is $\mu \notin P(\Theta)$, so that the model is misspecified. Let $\pi \in \mathcal{M}(\Theta)$ be an arbitrary prior distribution and $P_*\pi \in \mathcal{M}^2(X)$ its corresponding non-parametric prior. By removing an arbitrarily small amount of mass from $P_*\pi$ and placing it on μ one obtains an arbitrarily close prior distribution ν that is consistent with respect to the data generating distribution μ. Therefore although $P_*\pi$ and ν may be made arbitrarily close, their posterior distributions would remain asymptotically separated by a distance corresponding to the degree of misspecification of the model (the distance from μ to $P(\Theta)$).

2.4 Robustness and Computational Bayesian Inference

One of the reasons behind Bayesian inference's recent surge in popularity is the availability of computational methodologies and environments to compute the posteriors, such as Markov chain Monte Carlo (MCMC) simulations. Indeed, Diaconis' [20] "Markov chain Monte Carlo revolution", indicates that MCMC is a powerful tool to implement Bayesian inference. More generally, see Neal [69] for a review of the use of MCMC in inference.

If we consider the computation of Bayesian inference in the context of a *Machine Learning algorithm* then the practical effects of computation need to be incorporated; both the computational requirements and the impact on performance and qualitative robustness. Indeed, when posterior distributions are approximated using MCMC simulations, the robustness analysis naturally includes not only quantifying sensitivities with respect to the choice of prior but also the analysis of convergence and stability of the computational method. This is particularly true in Bayesian updating where Bayes' rule is applied iteratively and computed/approximated posterior distributions are then treated as prior distributions. The singular, and apparently antagonistic, relationship between *qualitative robustness* and *consistency* presented in this paper suggests that the metrics used to analyze convergence and *qualitative robustness* should be chosen with care and not independently from each other.

Originally, the convergence of MCMC has been generally analyzed in terms of total variation, see e.g. Roberts and Rosenthal [76, 77]. However, as cautioned by Gelman [35], although MCMC is a powerful tool for Bayesian inference, it is also "so easy to apply that there are risks of serious errors." One such risk appears to be related to the absence of analysis: According to Roberts and Rosenthal [76, Pg. 10] "rigorous quantitative bounds are not available in general" and more recently Madras and Sezer [64] assert that "quantitative rigorous results about realistic examples are scarce". Possibly to rectify this situation, recently the Wasserstein metric is also considered, see e.g. Gibbs [41] and Madras and Sezer [64]. For the definition see Dudley [26]. For a separable metric space, according to Zolotarev [93, Pg. 289], Szulga [83] gave the first proof of the Kantorovich-Rubinshtein theorem, see e.g. Dudley [26, Thm. 11.8.2], which asserts that the Wasserstein metric is equal to the Lipschitz metric on its domain. If moreover, S is a Polish metric space with diameter at most 1, then by Huber and Ronchetti [50,

Cor. 2.18] we have

$$d_{Pr}^2 \leq d_W \leq 2d_{Pr}$$

for all probability measures and, when the diameter is bounded, Gibbs and Su [42, Thm. 2] furthermore show that

$$d_{Pr}^2 \leq d_W \leq (diam(S) + 1)d_{Pr}\,.$$

Therefore, on Polish metric spaces of bounded diameter, the Wasserstein metric is equivalent to the Prokhorov metric and therefore metrizes weak convergence.

Now let us incorporate the MCMC analysis into the robustness analysis when the posteriors are computed using MCMC simulations. Let us further suppose that we are in a situation where we can prove when the computation will have converged, or prove that it has converged using diagnostics such as found in Carlin and Chib [13] and Cowles and Carlin [16]. It then appears that incorporating the MCMC simulation into the robustness analysis makes establishing robustness in any topology stronger than that which we can establish the convergence of the MCMC appear problematic. Consequently, it appears that establishing *qualitative robustness* of Bayesian inference in the total variation topology on $\mathcal{M}(\Theta)$ might be too much to ask, while weaker topologies, such as the Wasserstein or Prokhorov topologies, might facilitate practical robustness analysis for important problems.

3 Lorraine Schwartz' Theorem

As described in Section 2.2, robustness and consistency are closely related properties and consistency will be at the core of the mechanism generating non-robustness. The breakthrough in consistency for Bayesian inference is considered to be Schwartz' theorem [80, Thm. 6.1], so we use it as a model for consistency and the conditions sufficient to generate it. Stated in Barron, Schervish and Wasserman [3, Intro], Wasserman [88, Pg. 3] and Ghosal, Ghosh and Ramamoorthi [37, Cor. 1] for the nonparametric case, for the parametric case we will operate in *standard Borel spaces*; measurable spaces which are Borel isomorphic with a Borel subset of a Polish metric space. By Schervish [79, Thm. B.32], regular conditional probabilities exist for conditioning random variables with values in a standard Borel space. Moreover, when the parametric model is a Markov kernel and is dominated by a σ-finite measure, then by the Bayes' Theorem for densities Schervish [79, Thm. 1.31] we have, in addition, that the Bayes' rule for densities determines a valid family of densities for the regular conditional distributions. We say that a model $P : \Theta \to \mathcal{M}(X)$ is dominated if there exists a σ-finite Borel measure ν on X such that $P_\theta \ll \nu$, $\theta \in \Theta$.

Recall the Kullback-Leibler divergence K between two measures μ_1 and μ_2 defined by

$$K(\mu_1, \mu_2) := \int \log\left(\frac{d\mu_1}{d\nu} \Big/ \frac{d\mu_2}{d\nu}\right) d\mu_1\,,$$

where ν is any measure such that both μ_1 and μ_2 are absolutely continuous with respect to ν. It is well known that K is nonnegative, and that it is finite only if $\mu_1 \ll \mu_2$, and

in that case $K(\mu_1, \mu_2) = \int \log \frac{d\mu_1}{d\mu_2} d\mu_1$. From this we can define the Kullback-Leibler ball $K_\epsilon(\mu)$ of radius ϵ about $\mu \in \mathcal{M}(X)$ by $K_\epsilon(\mu) = \{\mu' \in \mathcal{M}(X) : K(\mu, \mu') \leq \epsilon\}$. For a model $P : \Theta \to \mathcal{M}(X)$, there is the pullback to a function K on Θ defined by $K(\theta_1, \theta_2) := K(P_{\theta_1}, P_{\theta_2})$ and when the model is dominated by a σ-finite measure ν, if we let $p(x|\theta) := \frac{dP_\theta}{d\nu}(x), x \in X$ be a realization of the Radon-Nikodym derivative, then the pullback has the form

$$K(\theta_1, \theta_2) := \int \log \frac{p(x|\theta_1)}{p(x|\theta_2)} dP_{\theta_1}(x) .$$

From this we define a Kullback-Leibler neighborhood of a point $\theta \in \Theta$ by

$$K_\epsilon(\theta) := \left\{\theta' \in \Theta : K(\theta, \theta') \leq \epsilon\right\} .$$

Let us define the set of priors $\mathcal{K}(\theta) \subset \mathcal{M}(\Theta)$ which have Kullback-Leibler *support at* θ by

$$\mathcal{K}(\theta) := \left\{\pi \in \mathcal{M}(\Theta) : \pi\big(K_\epsilon(\theta)\big) > 0, \quad \epsilon > 0\right\} ,$$

which implicitly requires that $K_\epsilon(\theta)$ be measurable[1] for all $\epsilon > 0$. Also let $\mathcal{K} \subset \mathcal{M}(\Theta)$ denote those measures with *global* Kullback-Leibler support, that is,

$$\mathcal{K} := \cap_{\theta \in \Theta} \mathcal{K}(\theta)$$

is the set of priors which have Kullback-Leibler support at all θ, and let $\mathcal{K}^{ae} \supset \mathcal{K}$, defined by

$$\mathcal{K}^{ae} := \left\{\pi \in \mathcal{M}(\Theta) : \pi\left\{\theta \in \Theta : \pi\big(K_\epsilon(\theta)\big) > 0, \, \epsilon > 0\right\} = 1\right\} , \tag{3.1}$$

denote the set of priors with *almost global* Kullback-Leibler support.

Let us address the measurability of the Kullback-Leibler neighborhoods $K_\epsilon(\theta) \subset \Theta$, $\epsilon > 0$. For the nonparametric case, Barron, Schervish and Wasserman [3, Lem. 11] demonstrate that the Kullback-Leibler neighborhoods $K_\epsilon(P_{\theta^*}) \subset \mathcal{M}(X)$ are measurable with respect to the strong topology restricted to the subspace of measures which are absolutely continuous with respect to a common σ-finite reference measure. For the parametric case, Dupuis and Ellis [27, Lem. 1.4.3] assert that on a Polish space that K is lower semicontinuous in both arguments. Since the subset embedding : $X \to X'$ of a subset X of a metric space X' is isometric, when X is a Borel subset of a separable metric space X', it can be shown that the induced pushforward map $i_* : \mathcal{M}(X) \to \mathcal{M}(X')$ is isometric in the Prokhorov metrics, in particular it is continuous. Since the composition of a continuous and a lower semicontinuous function is lower semicontinuous, it follows from Dupuis and Ellis [27, Lem. 1.4.3] that on any realization of a standard Borel space that the Kullback-Leibler divergence is lower semicontinuous in each of its

[1] Note the change from the standard definition $K_\epsilon(\mu) = \{\mu' : K(\mu, \mu') < \epsilon\}$ to ours $K_\epsilon(\mu) = \{\mu' : K(\mu, \mu') \leq \epsilon\}$ does not affect which measures have Kullback-Leibler support, but is more convenient since then $K_\epsilon(\mu)$ is closed, simplifying the proof that $K_\epsilon(\theta)$ is measurable.

arguments separately, in particular, fixing the first, it is lower semicontinuous. Therefore $K_\epsilon(P_{\theta^*}) \subset \mathcal{M}(X)$ is closed, and therefore measurable for $\epsilon > 0$. Consequently, when P is measurable, it follows that $K_\epsilon(\theta^*) \subset \Theta$ is measurable for $\epsilon > 0$.

The following corollary to Schwartz' Theorem, and its implications in Proposition 3.4, gives us the form of consistency that we will use in the robustness analysis. Since it assumes the model $P : \Theta \to \mathcal{M}(X)$ is measurable, and since by Aliprantis and Border [1, Thm. 15.13] the map $\mathcal{M}(X) \to \mathbb{R}$ defined by $\mu \mapsto \mu(A)$ is Borel measurable for all $A \in \mathcal{B}(X)$, it follows that P corresponds to a Markov kernel. Moreover since it also assumes the model to be dominated, it follows from Barron, Schervish and Wasserman [3, Lem. 10] that the Radon-Nikodym derivatives can be chosen so that they are $\mathcal{B}(X) \times \mathcal{B}(\Theta)$ measurable. Consequently, for a prior π, such a choice determines a well-defined conditional measure π_{x^n} for any n-sample $x^n \in X^n$. Note the assumption that the map $P : \Theta \to P(\Theta)$ be open.

Corollary 3.1 (Schwartz)**.** *Let X and Θ be Borel subsets of Polish metric spaces and equip $\mathcal{M}(X)$ and $\mathcal{M}(\Theta)$ with the Prokhorov metric. Consider an injective measurable dominated model $P : \Theta \to \mathcal{M}(X)$ with the family of conditional densities chosen to be $\mathcal{B}(X) \times \mathcal{B}(\Theta)$ measurable. Furthermore suppose that $P : \Theta \to P(\Theta)$ is an open map. Then for every $\pi \in \mathcal{M}(\Theta)$ with Kullback-Leibler support at $\theta^* \in \Theta$, for every measurable neighborhood U of θ^*, we have*

$$\pi_{x^n}(U) \to 1 \quad n \to \infty, \quad a.e.\ P_{\theta^*}^\infty .$$

Remark 3.2. Since Θ is a Borel subset of a Polish metric space and P is injective and measurable, it follows from Kechris' [55, Cor. 15.2] corollary to the Lusin-Souslin Theorem [55, Thm. 15.1], that $P(\Theta) \subset \mathcal{M}(X)$ is Borel. However, the additional assumption that $P : \Theta \to P(\Theta)$ be open is equivalent to assuming that $P^{-1} : P(\Theta) \to \Theta$ is continuous. In particular, it follows that $P^{-1} : P(\Theta) \to \Theta$ is measurable, so that $P : \Theta \to P(\Theta)$ is a Borel isomorphism.

Remark 3.3. When Θ and X are Borel subsets of Polish metric spaces, if P is injective and $P(\Theta) \subset \mathcal{M}(X)$ is discrete, in that every subset of $P(\Theta)$ is open in the relative topology, it follows that $P^{-1} : P(\Theta) \to \Theta$ is continuous and therefore $P : \Theta \to P(\Theta)$ is open. In addition, since separable discrete spaces are countable, see e.g. [82, Sec. II.3.8], it follows that $P(\Theta)$ is countable and therefore measurable. It also follows that Θ is countable, although it may not be discrete. Since $P^{-1}(A)$ is countable, and therefore measurable, for all measurable A, it follows that P is measurable. Consequently, in this case the measurability and openness conditions on the model P of Theorem 5.1 follow from the assumption that $P(\Theta) \subset \mathcal{M}(X)$ be discrete.

It will be useful to express the assertion of Corollary 3.1 and some of its consequences in terms of the convergence of measures and random measures. To that end, recall the notation $\mathcal{M}^2(\Theta) := \mathcal{M}(\mathcal{M}(\Theta))$, and consider the corresponding sequence of random variables $\pi_n : (X^\infty, P_{\theta^*}^\infty) \to \mathcal{M}(\Theta)$, defined by $\pi_n(x^\infty) := \pi_{x^n}$, $x^\infty \in X^\infty$, and its induced sequence of laws $(\pi_n)_* P_{\theta^*}^\infty \in \mathcal{M}^2(\Theta)$. Note especially that $\delta_{\delta_{\theta^*}}$ is the Dirac mass in $\mathcal{M}^2(\Theta)$ situated at the Dirac mass δ_{θ^*} in $\mathcal{M}(\Theta)$ situated at θ^*.

Proposition 3.4. *The assertion of Corollary 3.1 is equivalent to*

$$\pi_{x^n} \mapsto \delta_{\theta^*} \quad a.e.\ P_{\theta^*}^{\infty}\,,$$

where $\mapsto$ *is weak convergence. This in turn implies that*

$$P_{\theta^*}^{\infty}\Big\{d_{Pr}(\pi_n, \delta_{\theta^*}) > \epsilon\Big\} \to 0 \quad n \to \infty\,, \tag{3.2}$$

for $\epsilon > 0$*, which is equivalent to*

$$d_{Prr}\Big((\pi_n)_* P_{\theta^*}^{\infty}, \delta_{\delta_{\theta^*}}\Big) \to 0 \quad n \to \infty\,, \tag{3.3}$$

where d_{Prr} *is the Prokhorov metric on* $\mathcal{M}^2(\Theta)$ *defined with respect to the Prokhorov metric* d_{Pr} *on* $\mathcal{M}(\Theta)$*.*

4 Qualitative Robustness of Bayesian Inference

Hable and Christmann [45] have recently established *qualitative robustness* for support vector machines. Hampel [47] introduced the notion of the *qualitative robustness* of a sequence of estimators and Cuevas [18] has extended Hampel's definition and his basic structural results to Polish parameter spaces. Boente et al. [11] have developed *qualitative robustness* for stochastic processes and Nasser et al. [68] for estimation. The primary goal of this section is to develop a notion of *qualitative robustness* for Bayesian inference in the spirit of Hampel. To do so, in Section 4.1, we begin by demonstrating how Bayesian inference with a fixed prior can naturally be put into Hampel's framework, following Cuevas [18]. Then, in Section 4.2, we consider fixing the data generating distribution, and in Section 4.3 we combine the two into one coherent framework. Finally, in Section 4.4, we define a weaker form based on the Prokhorov metric on the space of measures on the space of measures equipped with the Prokhorov metric, and demonstrate how non-robustness with respect to this weaker form establishes non-robustness for the primary form. Under the assumptions of Schwartz' corollary, posterior distributions are well-defined for any multi-sample, so that for this discussion, we can disregard any well-definedness issues regarding the definition of the posterior and the resulting measure theoretic technicalities. For the more general case, to incorporate that conditioning is only defined almost everywhere, we refer to Mizera's [66] comprehensive extension of Hampel and Cuevas' results to multivalued mappings.

Of course, there are many variations on this theme, and the choice of metrics will affect not only the attainability of results, but the relevance of any results obtained. Metrics on spaces of measures is a well studied field, see e.g. Rachev et al. [75] and Gibbs and Su [42], but to keep the presentation simple, here we will restrict our attention to the total variation, Prokhorov and Ky Fan metrics. On general measurable spaces, we can metrize the space of measures $\mathcal{M}(X)$ and the space of measures on the space of measures $\mathcal{M}^2(\Theta)$ using total variation. However, when X is metric, we can also metrize the space of measures $\mathcal{M}(X)$ with the Prokhorov metric. In the same way, when Θ is

a metric space, we can metrize the space of measures $\mathcal{M}(\Theta)$ also with the Prokhorov metric, and having metrized in any way, we can then proceed to metrize $\mathcal{M}^2(\Theta)$ using the Prokhorov metric. Moreover, we remind the reader that, unlike Hampel and Cuevas who require "for all n" in their definitions, we follow Huber [50] and Mizera [66] in only requiring closeness "for large enough n". Finite sample versions, as introduced in Hable and Christmann [46, Def. 2], are also available.

4.1 Varying the data generating distribution and fixing the prior

Following Cuevas [18], we define qualitative robustness when varying the data generating distribution and fixing the prior.

Definition 4.1. Let $\mathcal{P} \subset \mathcal{M}(X)$ be an admissible set containing $\mu \in \mathcal{M}(X)$ and let $d_{\mathcal{M}(X)}$ and $d_{\mathcal{M}^2(\Theta)}$ be metrics on the spaces $\mathcal{M}(X)$ and $\mathcal{M}^2(\Theta)$ respectively. Then we say that the Bayesian inference for prior $\pi \in \mathcal{M}(\Theta)$ is qualitatively robust at μ, with respect to the subset $\mathcal{P}$ and the metrics $d_{\mathcal{M}(X)}$ and $d_{\mathcal{M}^2(\Theta)}$, if for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\acute{\mu} \in \mathcal{P},\ d_{\mathcal{M}(X)}(\mu, \acute{\mu}) < \delta \implies d_{\mathcal{M}^2(\Theta)}(\pi_* \mu^n, \pi_* \acute{\mu}^n) < \epsilon$$

for large enough n.

Using this setup with Polish spaces and the resulting Prokhorov metrics, Cuevas [18, Thm. 7] proves in two cases, that common Bayesian models with specific fixed priors are *not qualitatively robust.*

4.2 Varying the prior and fixing the data generating distribution

Regarding the importance of the robustness of Bayesian inference with respect to the prior, we quote from Berger's [7] discussion on Diaconis and Ylvisaker [22]:

> "There is a very serious issue concerning such an approximation, however, namely the issue of whether this good approximation to the prior ensures that the posterior will also be well approximated. I think the answer, in general, is no."

The stability of Bayesian decision theory with respect to the prior was fully initiated by Kadane and Chuang [52, 53] and further developed in Chuang [14] and Salinetti [78], and positive results obtained. In particular, in [52] comparison with previous notions, in particular that of Edwards, Lindman and Savage [29], is made. In Kadane and Srinivasan [51, 54] sufficient conditions for the stability of Bayes decision problems under uniform convergence of losses are obtained, generalizing the previously mentioned works [52, 14, 78].

However, we would like to proceed along the lines of Hampel's approach here, so that we can combine it with Section 4.1 to obtain a framework for *qualitative robustness* for Bayesian inference which simultaneously includes variation in the prior and the data

generating distribution into one Hampel-like framework in a natural way. This has the added advantage that we can utilize the fundamental results of Hampel [47] and Cuevas [18], and points to further development which will be useful. For example, Mizera [66] has fully developed these notions of *qualitative robustness* to include both ill-definedness and multi-valuedness, an extension which would be extremely useful for any treatment of Bayesian inference which incorporates conditioning only being determined almost everywhere. To that end, let us now develop a definition of *qualitative robustness* with respect to variation in the prior, for fixed data generating distribution, based on Hampel.

For a single sample $x \in X$, Basu et al.[5] say that the Bayesian inference is qualitatively robust at π and $x \in X$, with respect to a metric $d_{\mathcal{M}(\Theta)}$ on $\mathcal{M}(\Theta)$ and an admissible set $\Pi \subset \mathcal{M}(\Theta)$ containing π, if given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\acute{\pi} \in \Pi,\ d_{\mathcal{M}(\Theta)}(\pi, \acute{\pi}) < \delta \implies d_{\mathcal{M}(\Theta)}(\pi_x, \acute{\pi}_x) < \epsilon$$

They provide many positive results and some negative results regarding the Prokhorov and Levy metrics. We can extend this definition easily to a sequence $x^\infty := x_i, i = 1, ...,$ as follows: for the first n, we let $x^n := \{x_i, i = 1, .., n\}$ denote the n-sample and then define the sequence of posteriors $\pi_{x^n} \in \mathcal{M}(\Theta)$, $n = 1, ...$ Then we say that the Bayesian inference is qualitatively robust at π and $x^\infty \in X^\infty$ using the notion of stability of a dynamical system: if, given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\acute{\pi} \in \Pi,\ d_{\mathcal{M}(\Theta)}(\pi, \acute{\pi}) < \delta \implies d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) < \epsilon\,,$$

for large enough n. However, this definition puts no distributional requirements on the sequence x^∞. For an i.i.d. sample sequence $x^\infty \sim \mu^\infty$, we can include the "i.i.d. with respect to μ" assumption in the definition in a natural way by saying that the Bayesian inference is qualitatively robust at π if given $\epsilon_1, \epsilon_2 > 0$, there exists a $\delta > 0$ such that

$$\acute{\pi} \in \Pi,\ d_{\mathcal{M}(\Theta)}(\pi, \acute{\pi}) < \delta \implies \mu^n\big\{x^n : d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) > \epsilon_1\big\} < \epsilon_2$$

for large enough n. This definition can be calibrated with the following single parameter version: given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\acute{\pi} \in \Pi,\ d_{\mathcal{M}(\Theta)}(\pi, \acute{\pi}) < \delta \implies \mu^n\big\{x^n : d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) > \epsilon\big\} < \epsilon \tag{4.1}$$

for large enough n, where by "calibrated" we mean that if $\mu^n\big\{x^n : d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) > \epsilon_1\big\} < \epsilon_2$, then if we define $\epsilon := \max(\epsilon_1, \epsilon_2)$, we have

$$\begin{aligned} \mu^n\big\{x^n : d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) > \epsilon\big\} &\leq \mu^n\big\{x^n : d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) > \epsilon_1\big\} \\ &< \epsilon_2 \\ &< \epsilon\,, \end{aligned}$$

so that we conclude that $\mu^n\big\{x^n : d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) > \epsilon\big\} < \epsilon$, and conversely, if $\mu^n\big\{x^n : d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) > \epsilon\big\} < \epsilon$, then $\mu^n\big\{x^n : d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) > \epsilon_1\big\} < \epsilon_2$ with $\epsilon_1 := \epsilon$, $\epsilon_2 := \epsilon$.

We now express the condition (4.1) as convergence in probability of $\mathcal{M}(\Theta)$-valued random variables and metrize this convergence using the Ky Fan metric. To that end, consider the sequence of maps

$$\pi_n : X^\infty \to \mathcal{M}(\Theta)$$

defined by

$$\pi_n(x^\infty) := \pi_{x^n}, \quad x^\infty \in X^\infty .$$

Then for every $\mu \in \mathcal{M}(X)$, we use the same symbol π_n to denote the corresponding sequence

$$\pi_n : (X^\infty, \mu^\infty) \to \mathcal{M}(\Theta)$$

of $\mathcal{M}(\Theta)$-valued random variables defined on the probability space (X^∞, μ^∞).

To metrize the condition (4.1) as convergence of the random variables π_n using the Ky Fan metric, recall that for a metric spaces S, the metric $d : S \times S \to \mathbb{R}$ is a continuous function and therefore Borel measurable with respect to the Borel σ-algebra $\mathcal{B}(S \times S)$. However, in general, we have a proper inclusion $\mathcal{B}(S) \times \mathcal{B}(S) \subset \mathcal{B}(S \times S)$, so that, in general, the metric may not be measurable with respect to $\mathcal{B}(S) \times \mathcal{B}(S)$. However, when S is separable, it follows from Dudley [26, Prop. 4.1.7], that $\mathcal{B}(S) \times \mathcal{B}(S) = \mathcal{B}(S \times S)$, in which case we have the appropriate measurability of the metric function needed in the definition of the Ky Fan metric that will follow. This is the reason Rachev et al. [75, Rmk. 2.5.1] restricts attention to separable spaces. See Dudley [25] for the development of weak convergence of measures on nonseparable spaces.

For a separable metric space S, probability space (Ω, Σ, P), and two S-valued random variables $Z : \Omega \to S$ and $W : \Omega \to S$, the Ky Fan distance between Z and W, see e.g. [26, Pg. 289], is defined as

$$\alpha(Z, W) := \inf\{\epsilon \geq 0 : P(d(Z, W) > \epsilon) \leq \epsilon\} . \tag{4.2}$$

By Dudley [26, Thm. 9.2.2], the Ky Fan metric metrizes convergence in probability of S-valued random variables from (Ω, Σ, P).

To proceed, suppose that the metric space $\big(\mathcal{M}(\Theta), d_{\mathcal{M}(\Theta)}\big)$ is separable. In particular, note that when Θ is a separable metric space, such as a Borel subset of a Polish metric space, then the metric space $\big(\mathcal{M}(\Theta), d_{Pr}\big)$, where d_{Pr} is the Prokhorov metric, is separable. Then for fixed data generating measure μ and two priors $\pi, \acute{\pi} \in \mathcal{M}(\Theta)$, the identity

$$\begin{aligned} \mu^\infty\{d_{\mathcal{M}(\Theta)}(\pi_n, \acute{\pi}_n) > \epsilon\} &= \mu^\infty\{x^\infty : d_{\mathcal{M}(\Theta)}\big(\pi_n(x^\infty), \acute{\pi}_n(x^\infty)\big) > \epsilon\} \\ &= \mu^n\{x^n : d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) > \epsilon\} \end{aligned}$$

implies that the inequality

$$\mu^n\{x^n : d_{\mathcal{M}(\Theta)}(\pi_{x^n}, \acute{\pi}_{x^n}) > \epsilon\} < \epsilon$$

on the righthand side of (4.1) can be written

$$\mu^\infty\{d_{\mathcal{M}(\Theta)}(\pi_n, \acute{\pi}_n) > \epsilon\} < \epsilon$$

which, from the definition (4.2), implies that

$$\alpha_\mu(\pi_n, \acute{\pi}_n) \leq \epsilon\,,$$

where α_μ denotes the Ky Fan metric defined on the space of $\mathcal{M}(\Theta)$-valued random variables $W : (X^\infty, \mu^\infty) \to \mathcal{M}(\Theta)$ on the probability space (X^∞, μ^∞).

Since, for fixed data generating measure μ, the sequence of random variables $\pi_n, n = 1, ..$ are all defined on the same probability space (X^∞, μ^∞), the definition 4.1 of *qualitative robustness* now can be stated in terms of the Ky Fan metric on the space of $\mathcal{M}(\Theta)$-valued random variables.

Definition 4.2. Let $d_{\mathcal{M}(\Theta)}$ be a metric on $\mathcal{M}(\Theta)$ making it separable. Consider $\mu \in \mathcal{M}(X)$, $\pi \in \mathcal{M}(\Theta)$ and an admissible set Π containing π. Then the Bayesian inference is qualitatively robust if given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\acute{\pi} \in \Pi,\ d_{\mathcal{M}(\Theta)}(\pi, \acute{\pi}) < \delta \implies \alpha_\mu(\pi_n, \acute{\pi}_n) < \epsilon$$

for large enough n, where α_μ is the Ky Fan metric on the space of $\mathcal{M}(\Theta)$-valued random variables on the probability space (X^∞, μ^∞).

This seems to be the most reasonable generalization of Basu et al. [5] to i.i.d sequences. Of course, other definitions could be used -we simply must specify a metric α_μ on $\mathcal{M}(\Theta)$-valued random variables.

4.3 Qualitative Robustness of Bayesian Inference -Definition

Now we take the ideas of the previous two subsections and combine them to allow the variation in both the prior and the data generating distribution. Recall from Section 4.1 that when the prior π is fixed and we vary the data generating distribution μ, we define a map $\bar{\pi} : X^n \to \mathcal{M}(\Theta)$ by

$$\bar{\pi}(x^n) := \pi_{x^n}$$

and use the corresponding pushforward operator

$$\pi_* : \mathcal{M}(X^n) \to \mathcal{M}^2(\Theta)\,,$$

to pushforward μ^n to

$$\pi_*\mu^n \in \mathcal{M}^2(\Theta)\,.$$

Then we say that the Bayesian inference for prior $\pi \in \mathcal{M}(\Theta)$ is qualitatively robust at μ with respect to an admissible set $\mathcal{P}$ containing μ, and metrics $d_{\mathcal{M}(X)}$ and $d_{\mathcal{M}^2(\Theta)}$, if for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\acute{\mu} \in \mathcal{P},\ d_{\mathcal{M}(X)}(\mu, \acute{\mu}) < \delta \implies d_{\mathcal{M}^2(\Theta)}(\pi_*\mu^n, \pi_*\acute{\mu}^n) < \epsilon$$

for large enough n.

On the other hand, when the data generating distribution μ is fixed and we vary the prior π, we consider the sequence of maps

$$\pi_n : X^\infty \to \mathcal{M}(\Theta)$$

defined by

$$\pi_n(x^\infty) := \pi_{x^n}, \quad x^\infty \in X^\infty,$$

and the resulting sequence

$$\pi_n : (X^\infty, \mu^\infty) \to \mathcal{M}(\Theta)$$

of $\mathcal{M}(\Theta)$-valued random variable. Then, according to Definition 4.2, for an admissible set Π containing π, we say that the Bayesian inference is qualitatively robust if given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\acute{\pi} \in \Pi,\ d_{\mathcal{M}(\Theta)}(\pi, \acute{\pi}) < \delta \implies \alpha_\mu(\pi_n, \acute{\pi}_n) < \epsilon$$

for large enough n, where α_μ is the Ky Fan metric on the space of $\mathcal{M}(\Theta)$-valued random variables on the probability space (X^∞, μ^∞).

These two definitions can be combined in a straightforward manner to define robustness corresponding to a single prior/data generating pair. However, to consider a larger class of distributions than a single pair, we let $\mathcal{Z} \subset \big(\mathcal{M}(\Theta) \times \mathcal{M}(X)\big)^2$ denote the admissible set of prior-data generating distribution pairs $\big((\pi, \mu), (\acute{\pi}, \acute{\mu})\big) \in \big(\mathcal{M}(\Theta) \times \mathcal{M}(X)\big)^2$ such that $(\pi, \mu) \in \mathcal{M}(\Theta) \times \mathcal{M}(X)$ is an admissible candidate for robustness and $(\acute{\pi}, \acute{\mu}) \in \mathcal{M}(\Theta) \times \mathcal{M}(X)$ is an admissible candidate for its perturbation. In particular, the projection $\mathcal{Z}_1 \subset \mathcal{M}(\Theta) \times \mathcal{M}(X)$ denotes the set of admissible prior-data generating pairs. Now combining in a straightforward manner we obtain:

Definition 4.3. Consider a separable metric space $\big(\mathcal{M}(\Theta), d_{\mathcal{M}(\Theta)}\big)$ and metric spaces $\big(\mathcal{M}^2(\Theta), d_{\mathcal{M}^2(\Theta)}\big)$ and $\big(\mathcal{M}(X), d_{\mathcal{M}(X)}\big)$. For $\mu \in \mathcal{M}(X)$, let α_μ be a metric on the space of $\mathcal{M}(\Theta)$-valued random variables on the probability space (X^∞, μ^∞). Let $\mathcal{Z} \subset \big(\mathcal{M}(\Theta) \times \mathcal{M}(X)\big)^2$ denote the admissible set of prior-data generating distribution pairs. Then the Bayesian inference is qualitatively robust with respect to $\mathcal{Z}$, if given $\epsilon_1, \epsilon_2 > 0$, there exists $\delta_1, \delta_2 > 0$ such that

$$\big((\pi, \mu), (\acute{\pi}, \acute{\mu})\big) \in \mathcal{Z}, \quad d_{\mathcal{M}(\Theta)}(\pi, \acute{\pi}) < \delta_1,\ d_{\mathcal{M}(X)}(\mu, \acute{\mu}) < \delta_2$$
$$\implies d_{\mathcal{M}^2(\Theta)}(\acute{\pi}_* \mu^n, \acute{\pi}_* \acute{\mu}^n) < \epsilon_1, \quad \alpha_\mu(\pi_n, \acute{\pi}_n) < \epsilon_2$$

for large enough n.

4.4 Proprokhorov Robustness

Consider the Definition 4.3 of *qualitative robustness* using the Ky Fan metric. As mentioned before the Ky Fan metric only makes sense when the metric space $\mathcal{M}(\Theta)$ is separable. When Θ is is a separable metric space, such as a Borel subset of a Polish metric space, the space $\mathcal{M}(\Theta)$ equipped with the Prokhorov metric d_{Pr} is separable. Consequently we

now fix it to be $(\mathcal{M}(\Theta), d_{Pr})$. Let us call Definition 4.3 using the Ky Fan metric defined on the space of $(\mathcal{M}(\Theta), d_{Pr})$-valued random variables *Ky Fan-Prokhorov robustness*. We now define a weaker notion of *qualitative robustness* which we call *Proprokhorov robustness* such that Ky Fan-Prokhorov robustness implies Proprokhorov robustness. Consequently, and most importantly, Proprokhorov *non*-robustness implies Ky Fan-Prokhorov *non*-robustness. This weaker robustness has two distinct advantages. The first is that it has a simpler expression than Ky Fan-Prokhorov robustness and the second is that it is simpler to analyze. This simpler structure amounts to the Prokhorov metric on the space of probability measures on the space of probability measures equipped with the Prokhorov metric, suggesting the name Prokhorov-Prokhorov robustness, which we have shortened to Proprokhorov. Basic results which we will need regarding this metric space are derived in the appendix, Section 7.

To proceed, for fixed π and μ, the sequence

$$\pi_n : (X^\infty, \mu^\infty) \to \mathcal{M}(\Theta)\,, \quad n = 1, ...$$

of $\mathcal{M}(\Theta)$-valued random variables can be used to pushforward μ^∞ to the sequence

$$(\pi_n)_* \mu^\infty \in \mathcal{M}^2(\Theta)\,, \quad n = 1, ...$$

of laws in $\mathcal{M}^2(\Theta)$. Since, by definition $\pi_n(x^\infty) := \pi_{x^n}$ and the maps $\bar{\pi} : X^n \to \mathcal{M}(\Theta)$ were defined by $\bar{\pi}(x^n) := \pi_{x^n}$, dropping the $^-$ in the notation, it follows that

$$\pi_* \mu^n = (\pi_n)_* \mu^\infty, \quad n = 1, ... \tag{4.3}$$

According to Dudley [26, Thm. 11.3.5], for S-valued random variables Z, W from the same probability space, with laws μ_Z, μ_W, we have

$$d_{Pr}(\mu_Z, \mu_W) \leq \alpha(Z, W)\,. \tag{4.4}$$

Let us denote the Prokhorov metric on the space $\mathcal{M}^2(\Theta)$ by d_{Prr}. Then for fixed μ and priors π and $\acute{\pi}$, it follows from (4.3) and the Prokhorov-Ky Fan inequality (4.4) that

$$\begin{aligned} d_{Prr}(\pi_* \mu^n, \acute{\pi}_* \mu^n) &= d_{Prr}\big((\pi_n)_* \mu^\infty, (\acute{\pi}_n)_* \mu^\infty\big) \\ &\leq \alpha_\mu(\pi_n, \acute{\pi}_n) \end{aligned}$$

and so we conclude that

$$d_{Prr}(\pi_* \mu^n, \acute{\pi}_* \mu^n) \leq \alpha_\mu(\pi_n, \acute{\pi}_n)\,, \quad n = 1, ...\,.$$

From the triangle inequality we then obtain

$$\begin{aligned} d_{Prr}(\pi_* \mu^n, \acute{\pi}_* \acute{\mu}^n) &\leq d_{Prr}(\pi_* \mu^n, \acute{\pi}_* \mu^n) + d_{Prr}(\acute{\pi}_* \mu^n, \acute{\pi}_* \acute{\mu}^n) \\ &\leq \alpha_\mu(\pi_n, \acute{\pi}_n) + d_{Prr}(\acute{\pi}_* \mu^n, \acute{\pi}_* \acute{\mu}^n)\,, \end{aligned}$$

bounding the simple single term $d_{Prr}(\pi_* \mu^n, \acute{\pi}_* \acute{\mu}^n)$ in terms of the two terms $\alpha_\mu(\pi_n, \acute{\pi}_n)$ and $d_{Prr}(\acute{\pi}_* \mu^n, \acute{\pi}_* \acute{\mu}^n)$ of Ky Fan-Prokhorov robustness, defined in 4.3 using the Proprokhorov metric d_{Prr} on $\mathcal{M}^2(\Theta)$. Using the term $d_{Prr}(\pi_* \mu^n, \acute{\pi}_* \acute{\mu}^n)$ to define Proprokhorov robustness, it therefore follows that Ky Fan-Prokhorov robustness implies

Proprokhorov robustness articulated with respect to three parameters δ_1, δ_2 and ϵ and the Proprokhorov metric. By putting a metric on $\mathcal{M}(\Theta) \times \mathcal{M}(X)$ which is consistent with the product metric, we state an equivalent version in terms of two parameters δ and ϵ.

Definition 4.4 (Proprokhorov Robustness)**.** Let $\big(\mathcal{M}(X), d_{\mathcal{M}(X)}\big)$ be a metric space and let Θ be a separable metric space and consider the separable metric spaces $\big(\mathcal{M}(\Theta), d_{Pr}\big)$ and $\big(\mathcal{M}^2(\Theta), d_{Prr}\big)$. Let $\bar{d}$ be a metric which is consistent with the product metric $d_{Pr} \times d_{\mathcal{M}(X)}$ on $\mathcal{M}(\Theta) \times \mathcal{M}(X)$. Let $\mathcal{Z} \subset \big(\mathcal{M}(\Theta) \times \mathcal{M}(X)\big)^2$ denote the admissible set of prior-data generating distribution pairs. Then the Bayesian inference is qualitatively robust with respect to $\mathcal{Z}$, if given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\big((\pi, \mu), (\acute{\pi}, \acute{\mu})\big) \in \mathcal{Z},\ \bar{d}\big((\pi, \mu), (\acute{\pi}, \acute{\mu})\big) < \delta, \quad \Longrightarrow\ d_{Prr}(\pi_* \mu^n, \acute{\pi}_* \acute{\mu}^n) < \epsilon$$

for large enough n.

Since it is essential for our main results, we summarize the fact that Proprokhorov robustness is weaker than Ky Fan-Prokhorov robustness.

Theorem 4.5. *Let $\big(\mathcal{M}(X), d_{\mathcal{M}(X)}\big)$ be a metric space and let Θ be a separable metric space. Then Ky Fan-Prokhorov robustness of Definition 4.3, using the Ky Fan metric on the space of $(\mathcal{M}(\Theta), d_{Pr})$-valued random variables, implies Proprokhorov robustness of Definition 4.4.*

5 Main Results

Now that we have defined *qualitative robustness* for Bayesian inference and presented the consistency conditions of Section 3, we are now prepared for our main results. Indeed, the brittleness results of [71, 70, 72] and the non *qualitative robustness* results of Cuevas [18, Thm. 7] suggest that we may obtain non *qualitative robustness* according to Definition 4.3 by fixing the prior and varying the data generating distribution. However, according to Berk [9], in the misspecified case, although "there need be no convergence (in any sense)", in the limit the posterior becomes confined to a carrier set consisting of those points which are closest in terms of the Kullback-Leibler divergence. Consequently, it appears possible that a generalization of the results of Hampel [47, Lem. 3] and Cuevas [18, Thm. 1] which allows such a set-valued notion of consistency may be sufficient. Certainly it will require the more sophisticated notions of the continuity, or semi-continuity, of the Kullback-Leibler set-valued information projection and its dependence on the geometry of the model class $P(\Theta) \subset \mathcal{M}(X)$. Although this path will certainly be instructive and appears feasible, we instead find it simpler to obtain non *qualitative robustness* by fixing the data generating distribution to be in the model class and varying the prior. In particular, we show that the inference is not Proprokhorov robust according to Definition 4.4. It then follows from Theorem 4.5 that it is not Ky Fan-Prokhorov robust according to Definition 4.3, when the Ky Fan metric is defined on $(\mathcal{M}(\Theta), d_{Pr})$-valued random variables. It is important to note that these results do

not require any misspecification. Moreover, it appears that Bayesian Inference's dependence on both the data generating distribution and the prior leads to two complementary mechanisms generating non *qualitative robustness*; whereas Cuevas' result [18, Thm. 7] utilizes consistency and the discontinuity of the infinite sample limit, this other component utilizes *the non-robustness of consistency*, namely that the set of consistency priors, those with Kullback-Leibler support at the data generating distribution, is not robust.

Now let us return to our main results. For $\theta \in \Theta$, let us denote the set of priors with Kullback-Leibler support at θ by

$$\mathcal{K}(\theta) := \{\pi \in \mathcal{M}(\Theta) : \pi \text{ has Kullback-Leibler support at } \theta\}$$

and, for $\rho > 0$, define a *total variation* uniformity $\Pi_\rho(\theta) \subset \mathcal{M}(\Theta) \times \mathcal{M}(\Theta)$ by

$$\Pi_\rho(\theta) := \{(\pi, \acute{\pi}) \in \mathcal{M}(\Theta) \times \mathcal{M}(\Theta) : \pi \in \mathcal{K}(\theta),\ d_{tv}(\pi, \acute{\pi}) < \rho\}$$

of prior pairs where the first component has Kullback-Leibler support at θ and the second component is within ρ of the first in the total variation metric. For $\theta \in \Theta$, we define an admissible set of prior-data generating distribution pairs $\mathcal{Z}_\rho(\theta) \subset \left(\mathcal{M}(\Theta) \times \mathcal{M}(X)\right)^2$ by

$$\mathcal{Z}_\rho(\theta) := \Pi_\rho(\theta) \times P_\theta \times P_\theta\,, \tag{5.1}$$

using the identification of $\left(\mathcal{M}(\Theta) \times \mathcal{M}(X)\right)^2$ with $\mathcal{M}(\Theta) \times \mathcal{M}(\Theta) \times \mathcal{M}(X) \times \mathcal{M}(X)$.

Our Main Theorem shows, under the conditions of Schwartz' Corollary, that the Bayesian inference is not robust under the assumption that the prior has Kullback-Leibler support at the parameter value generating the data.

Theorem 5.1. *For all $\theta \in \Theta$, given the conditions of Schwartz' Corollary 3.1, the Bayesian inference is not Proprokhorov robust at $\mathcal{Z}_\rho(\theta)$ for all $\rho > 0$.*

Remark 5.2. Actually the proof shows more; let D denote the diameter of Θ, then for $\epsilon < \min\left(\frac{D}{2}, 1\right)$, there does not exist a $\delta > 0$ such that the Definition 4.4 of Proprokhorov robustness is satisfied. Since $\min\left(\frac{D}{2}, 1\right)$ is large, either half the diameter of the space or larger than 1, we say the inference is *brittle*.

Remark 5.3. In particular, Theorem 4.5 implies that the inference is not Ky Fan-Prokhorov robust.

Theorem 5.1 does not assert that the Bayesian inference is not robust at any specified prior, only that it is not robust under the assumption that the prior has Kullback-Leibler support at the parameter value generating the data. To establish non-robustness at specific priors we include variation in the data-generating distribution in the model class as follows. Let $\Delta_P \subset \mathcal{M}(X) \times \mathcal{M}(X)$, defined by

$$\Delta_P = \{(P_\theta, P_\theta), \theta \in \Theta\}\,,$$

denote the fact that we allow the data generating distribution to vary throughout the model class but do not allow any perturbations to it. Then, for $\pi \in \mathcal{M}(\Theta)$, define the admissible set $\mathcal{Z}_\rho(\pi) \subset \left(\mathcal{M}(\Theta) \times \mathcal{M}(X)\right)^2$ by

$$\mathcal{Z}_\rho(\pi) := \pi \times B_\rho^{tv}(\pi) \times \Delta_P\,,$$

where $B_\rho^{tv}(\pi)$ is the open ball in the total variation metric.

Since the following theorem is a corollary to the theorem after it, Theorem 5.5, we do not include its proof. However, we state it here because it is the more fundamental result.

Theorem 5.4. *Given the conditions of Schwartz' Corollary 3.1 with Θ not totally bounded. Then if the prior π has Kullback-Leibler support for all $\theta \in \Theta$, the Bayesian inference is not Proprokhorov robust at $\mathcal{Z}_\rho(\pi)$ for all $\rho > 0$.*

Since a metric space is totally bounded if and only if its completion is compact, when Θ is totally bounded, we assume that it is a Borel subset of a compact metric space. In this case, although Theorem 5.4 does not apply, utilizing the covering number and packing number inequalities of Kolmogorov and Tikhomirov [61], we can provide a natural *quantification of qualitative robustness*. To that end, we define covering and packing numbers. For a finite subset $\Theta' \subset \Theta$, the finite collection of open balls $\{B_\epsilon(\theta), \theta \in \Theta'\}$ is said to constitute a covering of Θ if $\Theta \subset \cup_{\theta\in\Theta'} B_\epsilon(\theta)$. For a finite set Θ' we denote its size by $|\Theta'|$. The covering numbers are defined by

$$\mathcal{N}_\epsilon(\Theta) = \min\left\{ |\Theta'| : \Theta \subset \cup_{\theta\in\Theta'} B_\epsilon(\theta) \right\},$$

that is, $\mathcal{N}_\epsilon(\Theta)$ is the smallest number of open balls of radius ϵ centered on points in Θ which covers Θ. On the other hand, a set of points $\Theta' \subset \Theta$ is said to constitute an ϵ-packing if $d(\theta_1, \theta_2) \geq \epsilon$, $\theta_1 \neq \theta_2 \in \Theta'$. The packing numbers are then defined by

$$\mathcal{M}_\epsilon(\Theta) := \max\left\{ |\Theta'| : \Theta' \text{ is an } \epsilon\text{-packing of } \Theta \right\}.$$

Since the Kolmogorov and Tikhomirov [61, Thm. IV] inequalities

$$\mathcal{M}_{2\epsilon}(\Theta) \leq \mathcal{N}_\epsilon(\Theta) \leq \mathcal{M}_\epsilon(\Theta) \tag{5.2}$$

are valid in the *not* totally bounded case, if we allow values of ∞, the following theorem has Theorem 5.4 as its corollary.

Theorem 5.5. *Given the conditions of Theorem 5.4 with Θ totally bounded and $\rho > 0$. If the Bayesian inference is Proprokhorov robust, then given $\epsilon > 0$, in terms of the product metric we must have*

$$\delta < \min\left(\frac{1}{\mathcal{N}_{2\epsilon}(\Theta)}, \rho \right).$$

Remark 5.6. Although the total variation metric on $\mathcal{M}(\Theta)$ is not used in the metric defining *qualitative robustness* for separability, measurability, and consistency purposes, the definition of the admissible sets $\mathcal{Z}_\rho(\theta)$ and $\mathcal{Z}_\rho(\pi)$ in terms of the total variation metric and a look at the proofs, can be used to show that these non *qualitative robustness* results primarily depend on total variation and do not arise because of the use of the weak topology.

6 Kullback-Leibler Support

Walker, Damien, and Lenk [87] argue that priors in Bayesian inference should be chosen to have Kullback-Leibler support for all $\theta \in \Theta$, that is they should have *global* Kullback-Leibler support. Moreover, the non-robustness mechanisms presented in Section 2.3 suggest the importance of selecting priors that are not "spread too thin". In this section we will discuss the relationship between these two notions under the condition that the model be measurable, injective, and open on its image in the weak topology.

Barron, Schervish and Wasserman [3], Petrone and Wasserman [73], Ghosal, Ghosh and Ramamoorthi [38], and Wu and Ghosal [92], demonstrate that priors with global Kullback-Leibler support exist in many important cases. However, in general, the existence of such priors appears to be nontrivial. Indeed recall that, just before Corollary 3.1, it was established that the measurability of the model in the weak topology implies the measurability of the Kullback-Leibler neighborhoods $\mathcal{K}_\epsilon(\theta)$ for $\theta \in \Theta$ and $\epsilon > 0$. Moreover, since the model is assumed to be open it follows that the image of any open set $\mathcal{O}$ in Θ is open in the relative weak topology of $P(\Theta) \subset \mathcal{M}(X)$ and therefore contains the intersection of $P(\Theta)$ with an open ball in $\mathcal{M}(X)$. By the inequality $d_{Pr} \leq d_{tv}$ and Pinsker's inequality $K \geq \frac{1}{2} d_{tv}^2$ we conclude that such an open ball contains a Kullback-Leibler neighborhood in $\mathcal{M}(X)$. Since P is injective, it follows that $\mathcal{O}$ contains its preimage, the corresponding Kullback-Leibler neighborhood in Θ. Since the assumption of global Kullback-Leibler support implies that the measure of this neighborhood is positive it follows that the measure of $\mathcal{O}$ is positive, and since $\mathcal{O}$ was arbitrary, we conclude that any measure with global Kullback-Leibler support is *strictly positive*, that is it satisfies Cromwell's rule in that the measure of every non-empty open set is positive. It is easy to show that a measure $\pi \in \mathcal{M}(\Theta)$ is strictly positive if and only if $\operatorname{supp} \pi = \Theta$.

However, the existence of strictly positive measures is also nontrivial and is connected with the Suslin Conjecture, see e.g. [56, 34]. Evidently, the foundations for this subject were developed in Kelley [56] and have been well-developed in Comfort and Negrepontis [15]. According to Argyros [2], the first example of a compact space satisfying the countable chain condition, i.e. such that every pairwise disjoint collection of non-empty open subsets is countable, and not carrying a strictly positive measure, was given by Gaifman [34]. Argyros [2] provides more examples under different conditions. Moreover, although it is well known that every compact topological group supports a strictly positive measure, Todorčevič [84] shows that the free topological group of the one-point compactification of the discrete space of size continuum does not support a strictly positive measure. On the other hand, Comfort and Negrepontis [15], van Casteren [85] and Plebanek [74], provide necessary and sufficient conditions for their existence. Finally, when Θ is a perfect (i.e. with no isolated points) compact metric space, Hebert and Lacey [48, Cor. 2.8] demonstrate that it possesses a continuous (i.e. vanishes on singletons) strictly positive measure. Therefore, we note that we have proven the the following lemma.

Lemma 6.1. *Let Θ and X be Borel subsets of Polish metric spaces, and suppose that the model $P : \Theta \to \mathcal{M}(X)$ is measurable, injective, and open on its image with respect*

to the weak topology. If Θ does not possess a strictly positive probability measure, then $\mathcal{K} = \emptyset$.

In the situation of Lemma 6.1, although Theorem 5.4 does not apply, an *almost everywhere* version of it can be proved under the the weaker assumption that π has *almost global* Kullback-Leibler support, that is $\pi \in \mathcal{K}^{ae}$ as defined in (3.1).

The following result illustrates density properties of the sets of measures with Kullback-Leibler support.

Lemma 6.2. *Let Θ and X be Borel subsets of Polish metric spaces and consider a measurable model $P : \Theta \to \mathcal{M}(X)$, where $\mathcal{M}(X)$ is equipped with weak topology, and the resulting Kullback-Leibler divergence K on $\Theta \times \Theta$. Then the set of probability measures $\mathcal{K}(\theta)$ with Kullback-Leibler support at θ is a convex dense subset of $\mathcal{M}(\Theta)$ in the total variation topology. Moreover, if the set of probability measures $\mathcal{K} := \cap_{\theta\in\Theta}\mathcal{K}(\theta)$ with Kullback-Leibler support at all $\theta \in \Theta$ is non-empty, it is a convex dense subset.*

7 Appendix: Some Prokhorov Geometry

We establish a basic mechanism to bound from below the Prokhorov distance between two measures based on the values of the measures on the neighborhood of a single set.

Lemma 7.1. *Let Z be a metric space and consider the space $\mathcal{M}(Z)$ of Borel probability measures equipped with the Prokhorov metric. Consider $\mu \in \mathcal{M}(Z)$ and suppose that there exists a set $B \in \mathcal{B}(Z)$ and $\alpha, \delta \geq 0$ such that*

$$\mu(B^\epsilon) \leq \delta, \quad \epsilon < \alpha\,.$$

Then, for any $\mu' \in \mathcal{M}(Z)$, we have

$$d_{Pr}(\mu, \mu') \geq \min\left(\alpha, \mu'(B) - \delta\right).$$

Proof. If $d_{Pr}(\mu_1, \mu_2) \geq \alpha$ the assertion is proved, so let us assume that $d_{Pr}(\mu_1, \mu_2) < \alpha$. Then, denoting $d^* := d_{Pr}(\mu_1, \mu_2)$, it follows from the assumption that $\mu(A^{d^*}) \leq \delta$, so that

$$\begin{aligned} \mu'(A) &\leq \mu(A^{d^*}) + d^* \\ &\leq \delta + d^* \end{aligned}$$

from which we conclude that $\mu'(A) - \delta \leq d^*$. Therefore, either $d_{Pr}(\mu_1, \mu_2) \geq \alpha$ or $d_{Pr}(\mu_1, \mu_2) \geq \mu'(A) - \delta$, proving the assertion. □

Lemma 7.2. *Let S be a separable metric space. Then, for an S-valued random variable X we have*

$$\alpha(X, s) = d_{Pr}(\mathcal{L}(X), \delta_s)$$

where α is the Ky Fan metric and s denotes the random variable with constant value s.

Proof. Let us denote $\alpha := \alpha(X, s)$ and $\rho := d_{Pr}(\mathcal{L}(X), \delta_s)$. Define the set $B_0 := \{s\}$ and $B_r := B_r(s), r > 0$ and observe that $B_0^r = B_r, r > 0$. Therefore, by the definition of ρ we have

$$\mathcal{L}(s)(B_0) \leq \mathcal{L}(X)(B_0^\rho) + \rho$$

and since $\mathcal{L}(s)(B_0) = 1$ we obtain

$$\mathcal{L}(X)(B_0^\rho) \geq 1 - \rho$$

from which we obtain $P(d(X, s) \geq \rho) \leq \rho$. Since this implies that

$$P(d(X, s) > \rho) \leq P(d(X, s) \geq \rho) \leq \rho$$

we conclude that $\rho \leq \alpha$. Since Dudley [26, Thm. 11.3.5] asserts that $\alpha \leq \rho$, the assertion follows. □

Proposition 7.3.

$$d_{Pr}(\delta_{x_1}, \delta_{x_2}) = \min\big(1, d(x_1, x_2)\big)$$

Proof. Consider the set $B := \{x_1\}$. Then since $B^\epsilon = B_\epsilon(x_1)$, it follows that for $\epsilon < d(x_1, x_2)$ that $x_2 \notin B^\epsilon$. Consequently, since $\delta_{x_1}(B) = 1$, the inequality

$$\delta_{x_1}(B) \leq \delta_{x_2}(B^\epsilon) + \epsilon$$

requires either $\epsilon \geq 1$ or $x_2 \in B^\epsilon$ which implies that $\epsilon \geq d(x_1, x_2)$. Consequently, $d_{Pr}(\delta_{x_1}, \delta_{x_2}) \geq \min\big(1, d(x_1, x_2)\big)$. To obtain equality, suppose that $d_{Pr}(\delta_{x_1}, \delta_{x_2}) > d(x_1, x_2)$. Then, for any d' which satisfies $d_{Pr}(\delta_{x_1}, \delta_{x_2}) > d' > d(x_1, x_2)$ there exists a measurable set B such that

$$\delta_{x_1}(B) > \delta_{x_2}(B^{d'}) + d'$$

Consequently, $x_1 \in B$, but $d' > d(x_1, x_2)$ implies that $x_2 \in B^{d'}$, which implies the contradiction $1 > 1 + d'$. □

8 Proofs

8.1 Proof of Corollary 3.1

We seek to apply Schwartz' theorem [80, Thm. 6.1]. Since U is a neighborhood it follows that it contains an open neighborhood O of θ. Since O is open and $P : \Theta \to P(\Theta)$ is open, it follows that $P(O)$ is open in $P(\Theta)$, and therefore there is an open set $V_* \subset \mathcal{M}(X)$ such that $V_* \cap P(\Theta) = P(O)$. Moreover, V_* is an open neighborhood of P_{θ^*}. Since X is a separable metric space, it follows that d_{Pr} metrizes the weak topology, and since V_* is open, it is well known (see e.g. [3, 88, 37]) that there exists a uniformly consistent test of P_{θ^*} against V_*^c, see Schwartz [80] for the definition of uniformly consistent test. It follows trivially that there exists a uniformly consistent test of P_{θ^*} against $V_*^c \cap P(\Theta)$.

Moreover, since P is injective it follows that $O^c = P^{-1}(V_*^c)$. Therefore, there exists a uniformly consistent test of P_{θ^*} against $V_*^c \cap P(\Theta) = \{P_\theta : \theta \in O^c\}$.

Since V_* is open, it also follows that there is a Prokhorov metric ball $B_s(P_{\theta^*})$ of radius $s > 0$ about P_{θ^*} such that $B_s(P_{\theta^*}) \subset V_*$. Now consider the Kullback-Leibler ball $K_\tau(P_{\theta^*})$ for $\tau < \frac{s^2}{2}$. It follows from Csiszar, Kemperman and Kullback's [17] improvement $K \geq \frac{1}{2} d_{tv}^2$ of Pinsker's inequality and the inequality $d_{tv} \geq d_{Pr}$, that $K_\tau(P_{\theta^*}) \subset B_s(P_{\theta^*})$. Since then $K_\tau(P_{\theta^*}) \subset B_s(P_{\theta^*}) \subset V_*$ it follows that

$$P^{-1}\big(K_\tau(P_{\theta^*})\big) \subset P^{-1}(V_*) = O\,.$$

Consider now the Kullback-Leibler neighborhood $W_\tau(\theta^*) \subset \Theta$ of θ^* defined by pulling $K_\tau(P_{\theta^*})$ back to Θ by the model P:

$$W_\tau(\theta^*) := P^{-1}\big(K_\tau(P_{\theta^*})\big)\,.$$

Then the previous inequality states that

$$W_\tau(\theta^*) \subset O\,.$$

Since the Kullback-Leibler neighborhoods are measurable in the weak topology and P is assumed measurable, it follows that $W_\tau(\theta^*)$ is measurable.

Therefore, O and $W_\tau(\theta^*)$ satisfy the assumptions of the sets V and and W in [80, Thm. 6.1]. Consequently, since by assumption, the prior π has Kullback-Leibler support, it follows that we can apply Schwartz' theorem [80, Thm. 6.1] to obtain the assertion for O and since $U \supset O$ is measurable the assertion follows.

8.2 Proof of Proposition 3.4

Let $\mathcal{O}$ denote the open sets in Θ and $\mathcal{O}_{\theta^*} \subset \mathcal{O}$ denote the open neighborhoods of θ^*. Then, under the conditions of Corollary 3.1, for $O \in \mathcal{O}_{\theta^*}$, it follows that

$$\pi_{x^n}(O) \to 1 \quad n \to \infty, \quad a.e.\ P_{\theta^*}^\infty\,.$$

Since $\delta_{\theta^*}(O) = 1$, $O \in \mathcal{O}_{\theta^*}$ and $\delta_{\theta^*}(O) = 0$, $O \in \mathcal{O} \setminus \mathcal{O}_{\theta^*}$ it easily follows that

$$\liminf_n \pi_{x^n}(O) \geq \delta_{\theta^*}(O),\ \forall O \in \mathcal{O}, \quad a.e.\ P_{\theta^*}^\infty\,.$$

which, by the Portmanteau theorem [26, Thm. 11.1.1], is equivalent to

$$\pi_{x^n} \mapsto \delta_{\theta^*} \quad a.e.\ P_{\theta^*}^\infty\,.$$

where $\mapsto$ denotes weak convergence.

Now consider the corresponding sequence of random variables $\pi_n : (X^\infty, P_{\theta^*}^\infty) \to \mathcal{M}(\Theta)$, defined by $\pi_n(x^\infty) := \pi_{x^n}$, $x^\infty \in X^\infty$, and its induced sequence of laws $(\pi_n)_* P_{\theta^*}^\infty \in \mathcal{M}^2(\Theta)$. Then $\pi_{x^n} \mapsto \delta_{\theta^*}$ *a.e.* $P_{\theta^*}^\infty$ is equivalent to

$$\pi_n \mapsto \delta_{\theta^*} \quad a.s.\ P_{\theta^*}^\infty\,.$$

Since Θ is a separable metric space it follows that $\mathcal{M}(\Theta)$ equipped with the Prokhorov metric is a separable metric space. Since a.s. convergence implies convergence in probability for random variables with values in a separable metric space, it follows that

$$\pi_n \mapsto \delta_{\theta^*} \text{ in } P_{\theta^*}^{\infty} - \text{probability}\,,$$

that is,

$$P_{\theta^*}^{\infty}\Big\{d_{Pr}(\pi_n, \delta_{\theta^*}) > \epsilon\Big\} \to 0 \quad n \to \infty\,.$$

Since $\mathcal{M}(\Theta)$ is a separable metric space it follows that $\mathcal{M}^2(\Theta)$ equipped with the Prokhorov metric is also a separable metric space. Therefore, since on separable metric spaces convergence in probability to a constant valued random variable is equivalent to the weak convergence of the corresponding set of laws to the Dirac mass situated at that value, see e.g. Dudley [26, Prop. 11.1.3], it follows that the convergence in probability, $\pi_n \to \delta_{\theta^*}$ in $P_{\theta^*}^{\infty}$ – probability, is equivalent to the corresponding convergence of laws

$$(\pi_n)_* P_{\theta^*}^{\infty} \mapsto \delta_{\delta_{\theta^*}} \quad n \to \infty\,.$$

Finally, since the Proprokhorov metric d_{Prr} on $\mathcal{M}^2(\Theta)$ metrizes the weak topology on $\mathcal{M}^2(\Theta) = \mathcal{M}(\mathcal{M}(\Theta))$, it follows that the latter is equivalent to

$$d_{Prr}\Big((\pi_n)_* P_{\theta^*}^{\infty}, \delta_{\delta_{\theta^*}}\Big) \to 0 \quad n \to \infty\,.$$

8.3 Proof of Theorem 5.1

Fix $\theta^* \in \Theta$, and consider another point $\theta \in \Theta$ and the Dirac mass $\delta_\theta \in \mathcal{M}(\Theta)$ situated at θ. From Lemma 6.2 we know that $\mathcal{K}(\theta^*)$ is dense in $\mathcal{M}(\Theta)$ in the total variation topology. In particular, for $\pi \in \mathcal{K}(\theta^*)$, the convex combination

$$\pi^\alpha := \alpha\pi + (1-\alpha)\delta_\theta$$

is a probability measure with Kullback-Leibler support, that is, $\pi^\alpha \in \mathcal{K}(\theta^*)$, $\alpha > 0$. and

$$d_{tv}(\pi^\alpha, \delta_\theta) \leq \alpha. \tag{8.1}$$

Therefore, it follows that

$$(\pi^\alpha, \delta_\theta) \in \Pi_\rho(\theta^*)\,, \quad \alpha < \rho\,,$$

and therefore

$$\big(\pi^\alpha, \delta_\theta, P_{\theta^*}, P_{\theta^*}\big) \in Z_\rho(\theta^*)\,, \quad \alpha < \rho\,,$$

where $\mathcal{Z}_\rho(\theta^*)$ is the admissible set defined in (5.1).

For the prior π^α, let $\pi_n^\alpha : (X^\infty, P_{\theta^*}^\infty) \to \mathcal{M}(\Theta)$, defined by $\pi_n^\alpha(x^\infty) := \pi_{x^n}^\alpha$, $x^\infty \in X^\infty$, denote the corresponding sequence of posterior random variables, and let $(\pi_n^\alpha)_* P_{\theta^*}^\infty \in \mathcal{M}^2(\Theta)$ denote its induced sequence of laws. On the other hand, for the prior δ_θ, it is easy to see that $(\delta_\theta)_{x^n} = \delta_\theta$, $x^n \in X^n$, so that if we denote the corresponding sequence of posterior random variables by δ_θ^n, then $(\delta_\theta^n)_* P_{\theta^*}^\infty = (\delta_\theta)_* P_{\theta^*}^n = \delta_{\delta_\theta}$.

Since the assumptions of Schwartz' Corollary 3.1 are satisfied and π^α has Kullback-Leibler support at θ^*, we can apply the assertion (3.2) of Proposition 3.4

$$P_{\theta^*}^\infty \Big\{ d_{Pr}(\pi_n^\alpha, \delta_{\theta^*}) > \epsilon \Big\} \to 0 \quad n \to \infty \,,$$

for $\epsilon > 0$. To complete the proof we simply use the fact that convergence in law to a Dirac mass is equivalent to convergence in probability to a constant random variable, that is use the equivalent assertion (3.3) of Proposition 3.4

$$d_{Prr}\Big((\pi_n^\alpha)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}} \Big) \to 0 \quad n \to \infty \,, \tag{8.2}$$

where d_{Prr} is the Prokhorov metric on $\mathcal{M}^2(\Theta)$. Now the proof is very simple. Indeed, from the triangle inequality we have

$$d_{Prr}\Big((\pi_n^\alpha)_* P_{\theta^*}^\infty, \delta_{\delta_\theta} \Big) \geq d_{Prr}\Big(\delta_{\delta_{\theta^*}}, \delta_{\delta_\theta} \Big) - d_{Prr}\Big((\pi_n^\alpha)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}} \Big)$$

and, by two applications of Proposition 7.3, we have

$$\begin{aligned} d_{Prr}\Big(\delta_{\delta_{\theta^*}}, \delta_{\delta_\theta} \Big) &= \min\Big(d_{Pr}\Big(\delta_{\theta^*}, \delta_\theta \Big), 1 \Big) \\ &= \min\Big(\min\big(d(\theta^*, \theta), 1 \big), 1 \Big) \\ &= \min\big(d(\theta^*, \theta), 1 \big) \,. \end{aligned}$$

Therefore, since $(\delta_\theta^n)_* P_{\theta^*}^\infty = \delta_{\delta_\theta}$, the convergence (8.2) implies that

$$d_{Prr}\Big((\pi_n^\alpha)_* P_{\theta^*}^\infty, (\delta_\theta^n)_* P_{\theta^*}^\infty \Big) \to \min\big(d(\theta^*, \theta), 1 \big) \,, \quad n \to \infty \,.$$

Finally, since $d_{Pr} \leq d_{tv}$, it follows from (8.1) that

$$d_{Pr}(\pi^\alpha, \delta_\theta) \leq \alpha.$$

Then, for any $\delta > 0$, if we restrict α so that $\alpha < \min(\delta, \rho)$, it follows that $d_{tv}(\pi^\alpha, \delta_\theta) < \rho$ and $d_{Pr}(\pi^\alpha, \delta_\theta) < \delta$, so that

$$\big(\pi^\alpha, \delta_\theta, P_{\theta^*}, P_{\theta^*} \big) \in \mathcal{Z}_\rho(\theta^*) \,, \tag{8.3}$$

$$d_{Pr}(\pi^\alpha, \delta_\theta) < \delta \,. \tag{8.4}$$

Let $D := \sup\{ d(\theta_1, \theta_2) : \theta_1, \theta_2 \in \Theta \}$ denote the diameter of Θ. Then it follows from the triangle inequality that, for any $\epsilon > 0$, there exists a $\theta \in \Theta$ such that $d(\theta^*, \theta) \geq \frac{D}{2} - \epsilon$. Consequently, for any $\bar{\epsilon} < \min(\frac{D}{2}, 1)$, no matter how small δ is, there is an $\alpha > 0$ such that, in addition to (8.3) and (8.4), we have

$$d_{Prr}\Big((\pi_n^\alpha)_* P_{\theta^*}^\infty, (\delta_\theta^n)_* P_{\theta^*}^\infty \Big) > \bar{\epsilon} \,,$$

for large enough n. Consequently, by the Definition 4.4, the Bayesian inference is not Proprokhorov robust.

8.4 Proof of Theorem 5.5

It follows from the definition of the packing numbers that, for $\epsilon > 0$, there is a packing $\{\theta_i, i = 1, .., \mathcal{M}_{2\epsilon}(\Theta)\}$ and therefore the collection of open balls $B_\epsilon(\theta_i)$, $i = 1, .., \mathcal{M}_{2\epsilon}(\Theta)$ is a disjoint union. Denoting $\mathcal{N}_{2\epsilon} := \mathcal{N}_{2\epsilon}(\Theta)$ and $\mathcal{M}_{2\epsilon} := \mathcal{M}_{2\epsilon}(\Theta)$, we therefore obtain

$$\begin{aligned} 1 &= \pi(\Theta) \\ &\geq \pi\big(\cup_{i=1}^{\mathcal{M}_{2\epsilon}} B_\epsilon(\theta_i)\big) \\ &= \sum_{i=1}^{\mathcal{M}_{2\epsilon}} \pi\big(B_\epsilon(\theta_i)\big) \\ &\geq \mathcal{M}_{2\epsilon} \min_{i=1,\mathcal{M}_{2\epsilon}} \pi\big(B_\epsilon(\theta_i)\big)\,. \end{aligned}$$

Consequently, since (5.2) implies $\mathcal{M}_{2\epsilon} \geq \mathcal{N}_{2\epsilon}$, there exists a point $\theta^* \in \Theta$ such that

$$\pi\big(B_\epsilon(\theta^*)\big) \leq \frac{1}{\mathcal{N}_{2\epsilon}}\,. \tag{8.5}$$

Let $B_\epsilon := B_\epsilon(\theta^*)$ denote the open ball about θ^* and let B_ϵ^c denote its complement. Let $\pi^\epsilon \in \mathcal{M}(\Theta)$, defined by

$$\pi^\epsilon(B) := \frac{\pi(B_\epsilon^c \cap B)}{\pi(B_\epsilon^c)},\ B \in \mathcal{B}(\Theta)\,,$$

denote the normalization of the restriction of π to B_ϵ^c which, by the inequality (8.5), is well defined. Since $\pi = \pi(B_\epsilon^c)\pi^\epsilon + \pi|_{B_\epsilon}$ it follows that $\pi - \pi^\epsilon = \pi|_{B_\epsilon} - \pi(B_\epsilon)\pi^\epsilon$ so that we obtain

$$d_{tv}(\pi^\epsilon, \pi) \leq \pi(B_\epsilon) \leq \frac{1}{\mathcal{N}_{2\epsilon}}$$

from which we obtain

$$d_{Pr}(\pi^\epsilon, \pi) \leq \frac{1}{\mathcal{N}_{2\epsilon}}\,. \tag{8.6}$$

In particular, when $\frac{1}{\mathcal{N}_{2\epsilon}} < \rho$, we obtain

$$\pi^\epsilon \in B_\rho^{tv}(\pi)$$

and therefore

$$\big(\pi, \pi^\epsilon, P_{\theta^*}, P_{\theta^*}\big) \in Z_\rho(\pi)\,.$$

That is, when $\frac{1}{\mathcal{N}_{2\epsilon}} < \rho$, the point $\big(\pi, \pi^\epsilon, P_{\theta^*}, P_{\theta^*}\big) \in Z_\rho(\pi)$.

For the prior π^ϵ, let $\pi_n^\epsilon : (X^\infty, P_{\theta^*}^\infty) \to \mathcal{M}(\Theta)$, defined by $\pi_n^\epsilon(x^\infty) := \pi_{x^n}^\epsilon$, $x^\infty \in X^\infty$, denote the corresponding sequence of posterior random variables, and let $(\pi_n^\epsilon)_* P_{\theta^*}^\infty \in \mathcal{M}^2(\Theta)$ denote its induced sequence of laws. Since the assumptions of Schwartz' Corollary 3.1 are satisfied and π has Kullback-Leibler support at θ^*, we can apply the assertion (3.3) of Proposition 3.4 to the sequence of posterior laws $(\pi_n)_* P_{\theta^*}^\infty$ corresponding to π:

$$d_{Prr}\Big((\pi_n)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}}\Big) \to 0 \quad n \to \infty\,. \tag{8.7}$$

From the triangle inequality we have

$$d_{Prr}\Big((\pi_n)_* P^\infty_{\theta^*}, (\pi^\epsilon_n)_* P^\infty_{\theta^*}\Big) \geq d_{Prr}\Big((\pi^\epsilon_n)_* P^\infty_{\theta^*}, \delta_{\delta_{\theta^*}}\Big) - d_{Prr}\Big((\pi_n)_* P^\infty_{\theta^*}, \delta_{\delta_{\theta^*}}\Big)\,, \tag{8.8}$$

so to lower bound the lefthand side it is sufficient in the limit to lower bound the first term on the right. To that end, we use a quantitative version of the partial converse [26, Thm. 11.3.5] of convergence in probability implies convergence in law, valid when the convergence in law is to a Dirac mass. Indeed, if we denote the Ky Fan metric determined from the measure $P^\infty_{\theta^*}$ by α_{θ^*}, Lemma 7.2 asserts that

$$d_{Prr}\Big((\pi^\epsilon_n)_* P^\infty_{\theta^*}, \delta_{\delta_{\theta^*}}\Big) = \alpha_{\theta^*}(\pi^\epsilon_n, \delta_{\theta^*})\,. \tag{8.9}$$

To evaluate the Ky Fan distance on the righthand side, first observe that since π^ϵ has support contained in the closed set B^c_ϵ, it follows from Schervish [79, Thm. 1.31] that $\pi^\epsilon_{x^n}$ also has support contained in B^c_ϵ a.e $P^n_{\theta^*}$. Therefore, if we define $B_0 := \{\theta^*\}$ and $B_r := B_r(\theta^*)$, it follows that $B^r_0 = B_r$, so that

$$\pi^\epsilon_{x^n}(B^r_0) = 0\,, \quad a.e.\ P^n_{\theta^*}, \quad r < \epsilon$$

and

$$(\delta_{\theta^*})_{x^n}(B_0) = 1\,, \quad a.e.\ P^n_{\theta^*}\,.$$

It follows from Lemma 7.1 that

$$d_{Pr}\big(\pi^\epsilon_{x^n}, (\delta_{\theta^*})_{x^n}\big) \quad \geq \quad \min(\epsilon, 1) \quad a.e.\ P^\infty_{\theta^*}\,,$$

and, since $\epsilon \leq 1$, we obtain

$$P^\infty_{\theta^*}\Big(d_{Pr}\big(\pi^\epsilon_{x^n}, (\delta_{\theta^*})_{x^n}\big) \geq \epsilon\Big) = 1\,.$$

Therefore, by the definition (4.2) of the Ky Fan metric, we obtain $\alpha_{\theta^*}(\pi^\epsilon_n, \delta_{\theta^*}) \geq \epsilon$ and, by the identity (8.9), we conclude that

$$d_{Prr}\Big((\pi^\epsilon_n)_* P^\infty_{\theta^*}, \delta_{\delta_{\theta^*}}\Big) \geq \epsilon\,.$$

Consequently, from the triangle inequality (8.8) and the convergence (8.7), we conclude, for any $\acute{\epsilon} > 0$, that for large enough n we have

$$d_{Prr}\Big((\pi_n)_* P^\infty_{\theta^*}, (\pi^\epsilon_n)_* P^\infty_{\theta^*}\Big) \geq \epsilon - \acute{\epsilon}\,. \tag{8.10}$$

Consequently, if this Bayesian inference is Proprokhorov robust, then for $\epsilon > 0$, it follows from (8.10) and (8.6) that $\delta < \frac{1}{\mathcal{N}_{2\epsilon}}$. The requirement that perturbations be admissible, that is determine members in $\mathcal{Z}_\rho(\pi)$, implies that $\delta < \rho$.

8.5 Proof of Lemma 6.2

The condition that $\pi \in \mathcal{M}(\Theta)$ have Kullback-Leibler support at θ, that is $\pi \in \mathcal{K}(\theta)$, is both projective and monotonic in the following sense. It is projective in that if $\pi \in \mathcal{K}(\theta)$ then $\alpha\pi \in \mathcal{K}(\theta)$ for $\alpha > 0$, and it is monotonic in the sense that $\pi' \geq \pi$ and $\pi \in \mathcal{K}(\theta)$ implies that $\pi' \in \mathcal{K}(\theta)$. The same is true for the condition $\pi \in \mathcal{K}$. Consequently, consider $\pi \in \mathcal{K}(\theta)$ and consider any $\acute{\pi} \in \mathcal{M}(\Theta)$. Then it follows that $\pi_\alpha := \alpha\pi + (1-\alpha)\acute{\pi} \in \mathcal{K}(\theta)$ for all $\alpha > 0$. Since

$$\pi_\alpha - \acute{\pi} = \alpha(\pi - \acute{\pi})$$

it follows that $d_{tv}(\pi_\alpha, \acute{\pi}) \leq \alpha$ and since $\alpha > 0$ was arbitrary the result is proved. The proof is the same for $\mathcal{K}$.

Acknowledgments

The authors gratefully acknowledge this work supported by the Air Force Office of Scientific Research under Award Number FA9550-12-1-0389 (Scientific Computation of Optimal Statistical Estimators).

References

[1] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, Berlin, third edition, 2006.

[2] S. Argyros. On compact spaces without strictly positive measure. *Pacific Journal of Mathematics*, 105(2):257–272, 1983.

[3] A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561, 1999.

[4] A. R. Barron. Discussion: On the consistency of Bayes estimates. *The Annals of statistics*, pages 26–30, 1986.

[5] S. Basu, S. R. Jammalamadaka, and W. Liu. Stability and infinitesimal robustness of posterior distributions and posterior quantities. *Journal of statistical planning and inference*, 71(1):151–162, 1998.

[6] J. O. Berger. The robust Bayesian viewpoint. In *Robustness of Bayesian Analyses*, volume 4 of *Stud. Bayesian Econometrics*, pages 63–144. North-Holland, Amsterdam, 1984. With comments and with a reply by the author.

[7] J. O. Berger. Discussion of "Quantifying prior opinion" by Diaconis and Ylvisaker. *Bayesian Statistics*, 2:133–156, 1985.

[8] J. O. Berger. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994. With comments and a rejoinder by the author.

[9] R. H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist. 37 (1966), 51–58; correction, ibid*, 37:745–746, 1966.

[10] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, pages 882–886, 1962.

[11] G. Boente, R. Fraiman, and V. J. Yohai. Qualitative robustness for stochastic processes. *The Annals of Statistics*, pages 1293–1312, 1987.

[12] G. E. P. Box. Non-normality and tests on variances. *Biometrika*, 40:318–335, 1953.

[13] B. P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 473–484, 1995.

[14] D. T. Chuang. Further theory of stable decisions. In *Robustness of Bayesian Analyses*, volume 4 of *Stud. Bayesian Econometrics*, pages 165–228. North-Holland, Amsterdam, 1984.

[15] W. W. Comfort and S. Negrepontis. *Chain Conditions in Topology*, volume 79. Cambridge University Press Cambridge-New York, 1982.

[16] M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[17] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975.

[18] A. Cuevas. Qualitative robustness in abstract inference. *Journal of statistical planning and inference*, 18(3):277–289, 1988.

[19] A. K. Dey and F. H. Ruymgaart. Direct density estimation as an ill-posed inverse estimation problem. *Statistica Neerlandica*, 53(3):309–326, 1999.

[20] P. Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

[21] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *Ann. Statist.*, 14(1):1–67, 1986. With a discussion and a rejoinder by the authors.

[22] P. Diaconis and D. Ylvisaker. Quantifying prior opinion. *Bayesian statistics*, 2:133–156, 1985.

[23] J. L. Doob. Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pages 23–27. Centre National de la Recherche Scientifique, Paris, 1949.

[24] C. J. Douady, F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. P. Douzery. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, 20(2):248–254, 2003.

[25] R. M. Dudley. Weak convergence of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces. *Illinois Journal of Mathematics*, 10(1):109–126, 1966.

[26] R. M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[27] P. Dupuis and R. S. Ellis. *A weak convergence approach to the theory of large deviations*, volume 902. John Wiley & Sons, 2011.

[28] A. W. F. Edwards. *Likelihood*. Johns Hopkins University Press, Baltimore, expanded edition, 1992.

[29] W. Edwards, H. Lindman, and L. J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193, 1963.

[30] D. Freedman. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics*, pages 1119–1140, 1999.

[31] D. Freedman and P. Diaconis. On inconsistent Bayes estimates in the discrete case. *The Annals of Statistics*, pages 1109–1118, 1983.

[32] D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pages 1386–1403, 1963.

[33] D. A. Freedman. On the asymptotic behavior of bayes estimates in the discrete case II. *The Annals of Mathematical Statistics*, pages 454–456, 1965.

[34] H. Gaifman. Concerning measures on Boolean algebras. *Pacific J. Math*, 14:61–73, 1964.

[35] A. Gelman. Inference and monitoring convergence. In *Markov chain Monte Carlo in practice*, pages 131–143. Springer, 1996.

[36] S. Ghosal. A review of consistency and convergence of posterior distribution. In *Varanashi Symposium in Bayesian Inference, Banaras Hindu University*, 1997.

[37] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Consistency issues in Bayesian nonparametrics. *Statistics Textbooks and Monographs*, 158:639–668, 1999.

[38] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *Annals of Statistics*, 27(1):143–158, 1999.

[39] S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

[40] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics. Springer-Verlag, New York, 2003.

[41] A. L. Gibbs. Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stochastic models*, 20(4):473–492, 2004.

[42] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[43] P. Grünwald and J. Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2-3):119–149, 2007.

[44] P. D. Grünwald. Bayesian inconsistency under misspecification. 2006. http://homepages.cwi.nl/~pdg/ftp/valenciapost.pdf.

[45] R. Hable and A. Christmann. On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, 102(6):993–1007, 2011.

[46] R. Hable and A. Christmann. Robustness versus consistency in ill-posed classification and regression problems. In *Classification and Data Mining*, pages 27–35. Springer, 2013.

[47] F. R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896, 1971.

[48] D. Hebert and H. Lacey. On supports of regular Borel measures. *Pacific Journal of Mathematics*, 27(1):101–118, 1968.

[49] T.-M. Huang. Convergence rates for posterior distributions and adaptive estimation. *The Annals of Statistics*, 32(4):1556–1593, 2004.

[50] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition, 2009.

[51] J Kadane and C. Srinivasan. Bayesian robustness and stability [with discussion and rejoinder]. *Lecture Notes-Monograph Series*, pages 81–100, 1996.

[52] J. B. Kadane and D. T. Chuang. Stable decision problems. *The Annals of Statistics*, pages 1095–1110, 1978.

[53] J. B. Kadane and D. T. Chuang. Stable decision problems. In *Robustness of Bayesian Analyses*, volume 4 of *Stud. Bayesian Econometrics*, pages 145–164. North-Holland, Amsterdam, 1984.

[54] J. B. Kadane and C. Srinivasan. Bayes decision problems and stability. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 383–404, 1998.

[55] A. S. Kechris. *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. Springer-Verlag, New York, 1995.

[56] J. L. Kelley. Measures on Boolean algebras. *Pacific J. Math*, 9(11):1165–1177, 1959.

[57] B. J. K. Kleijn. Bayesian asymptotics under misspecification. 2004. http://dspace.ubvu.vu.nl/bitstream/handle/1871/10842/6757.pdf?sequence=1.

[58] B. J. K. Kleijn. Criteria for Bayesian consistency. *arXiv preprint arXiv:1308.1263*, 2013.

[59] B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2):837–877, 2006.

[60] B. J. K. Kleijn and A. W. van der Vaart. The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Stat.*, 6:354–381, 2012.

[61] A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Trans. Amer. Math. Soc*, 17:277–364, 1961.

[62] L. Le Cam and L. Schwartz. A necessary and sufficient condition for the existence of consistent estimates. *The Annals of Mathematical Statistics*, pages 140–150, 1960.

[63] T. Lubik and F. Schorfheide. A Bayesian look at the new open economy macroeconomics. In *NBER Macroeconomics Annual 2005, Volume 20*, pages 313–382. MIT Press, 2006.

[64] N. Madras and D. Sezer. Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli*, 16(3):882–908, 2010.

[65] R. Martin, L. Hong, and S. G. Walker. A note on Bayesian convergence rates under local prior support conditions. *arXiv preprint arXiv:1201.3102*, 2012.

[66] I. Mizera. Qualitative robustness and weak continuity: the extreme unction. *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurecková*, 1:169, 2010.

[67] U. K. Müller. Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849, 2013.

[68] M. Nasser, N. A. Hamzah, and Md. A. Alam. Qualitative robustness in estimation. *Pakistan Journal of Statistics and Operation Research*, 8(3):619–634, 2012.

[69] R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. 1993. https://www.cs.princeton.edu/courses/archive/fall07/cos597C/readings/Neal1993.pdf.

[70] H. Owhadi and C. Scovel. Brittleness of Bayesian inference and new Selberg formulas. 2013. http://arxiv.org/abs/1304.7046v2.

[71] H. Owhadi, C. Scovel, and T. J. Sullivan. Bayesian Brittleness. 2013. http://arxiv.org/abs/1304.6772v2.

[72] H. Owhadi, C. Scovel, and T. J. Sullivan. On the Brittleness of Bayesian Inference. 2013. http://arxiv.org/abs/1308.6306v2.

[73] S. Petrone and L. Wasserman. Consistency of Bernstein polynomial posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(1):79–100, 2002.

[74] G. Plebanek. A note on strictly positive Radon measures. In *Colloq. Math*, volume 69, pages 187–192, 1995.

[75] S. T. Rachev, L. B. Klebakov, S. V. Stoyanov, and F. J. Fabozzi. *The Methods of Distances in the Theory of Probability and Statistics*. Springer, New York, 2013.

[76] G. O. Roberts and J. S. Rosenthal. Markov-chain Monte Carlo: Some practical implications of theoretical results. *Canadian Journal of Statistics*, 26(1):5–20, 1998.

[77] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[78] G. Salinetti. Stability of Bayesian decisions. *Journal of Statistical Planning and Inference*, 40(2):313–329, 1994.

[79] M. J. Schervish. *Theory of Statistics*. Springer, 1995.

[80] L. Schwartz. On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 4:10–26, 1965.

[81] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714, 2001.

[82] L. A. Steen and J. A. Seebach Jr. *Counterexamples in Topology*. Holt, Rinehart and Winston, 1970.

[83] A. Szulga. On minimal metrics in the space of random variables. *Theory of Probability & Its Applications*, 27(2):424–430, 1983.

[84] S. Todorčevič. Some applications of S and L combinatorics. *Annals of the New York Academy of Sciences*, 705(1):130–167, 1993.

[85] J. A. Van Casteren. Strictly positive Radon measures. *Journal of the London Mathematical Society*, 49(1):109–123, 1994.

[86] S. Walker. New approaches to Bayesian consistency. *Annals of Statistics*, pages 2028–2043, 2004.

[87] S. Walker, P. Damien, and P. Lenk. On priors with a Kullback–Leibler property. *Journal of the American Statistical Association*, 99(466), 2004.

[88] L. Wasserman. Asymptotic properties of nonparametric Bayesian procedures. In *Practical nonparametric and semiparametric Bayesian statistics*, pages 293–304. Springer, 1998.

[89] L. Wasserman, M. Lavine, and R. L. Wolpert. Linearization of Bayesian robustness problems. *J. Statist. Plann. Inference*, 37(3):307–316, 1993.

[90] L. Wasserman and T. Seidenfeld. The dilation phenomenon in robust Bayesian inference. *J. Statist. Plann. Inference*, 40:345–356, 1994.

[91] L. A. Wasserman. Prior envelopes based on belief functions. *Ann. Statist.*, 18(1):454–464, 1990.

[92] Y. Wu and S. Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, 2:298–331, 2008.

[93] V. M. Zolotarev. Probability metrics. *Theory of Probability & Its Applications*, 28(2):278–302, 1984.